SOLID-STATE ELECTRONICS

AN INTERNATIONAL JOURNAL

VOLUME 3, 1961

Editor-in-Chief
W. CRAWFORD DUNLAP
Waltham, Mass.

Regional Editors

U.K.
L. PINCHERLE
London

EUROPE
H. WELKER
Erlangen, Germany

JAPAN and FAR EAST
G. M. HATOYAMA
Yokohama

PERGAMON PRESS
OXFORD · LONDON · NEW YORK · PARIS

SOLID-STATE ELECTRONICS

AN INTERNATIONAL JOURNAL

VOLUME 3, 1961

Editor-in-Chief
W. CRAWFORD DUNLAP
Waltham, Mass.

Regional Editors

U.K.
L. PINCHERLE
London

EUROPE
H. WELKER
Erlangen, Germany

JAPAN and FAR EAST
G. M. HATOYAMA
Yokohama

PERGAMON PRESS
OXFORD · LONDON · NEW YORK · PARIS

LIST OF CONTENTS

NUMBER 1

M. Green, I. N. Greenberg, J. Banzhaf, A. Sass, H. S. White and R. V. McKnight: Electrical characteristics of a reduced-area, rhodium-plated, silicon end contact 1
W. T. Eriksen and H. Ahlburg: An analysis of electroluminescent brightness waves . . 11
C. Fa and R. Zuleeg: Field-retarded diffusion of antimony in germanium 18
D. F. Edwards, C. D. Butter and L. D. McGlauchlin: Photoconductivity in single-crystal tellurium 24
Kogo Sato and Masami Tomono: Current dependence of α_{cb} in the alloy junction transistor . 29
D. T. J. Hurle: Constitutional supercooling during crystal growth from stirred melts.—I. Theoretical 37
Kurt Lehovec and Alexis Slobodskoy: Diffusion of charged particles into a semiconductor under consideration of the built-in field 45
W. Heywang: Bariumtitanat als Sperrschichthalbleiter 51
Wen-Hsuing Ko: The forward transient behavior of semiconductor junction diodes . . 59

Notes:
T. C. Taylor and F. L. Yuan: Ge or Si junction made by the surface weld-regrowth process. 70
C. H. L. Goodman: Melting phenomena with silicon 72
P. Lloyd: Thin-film capacitors using tantalum oxide dielectrics prepared by reactive sputtering 74

Book Reviews:
Photoconductivity of Solids, by Richard H. Bube (Reviewed by R. W. Redington) . . 76
The Hall Effect and Related Phenomena, by E. H. Putley (Reviewed by W. C. Dunlap) . 76
High Frequency Applications of Ferrites, by J. Roberts (Reviewed by Jerome J. Green) . 77

NUMBER 2

I. Braun and E. Y. Wang: A new method of zone leveling for materials having distribution coefficients in the range of 0·0·6 79
J. E. Parrott and A. W. Penn: The design theory of thermoelectric cooling elements and units . 91
F. G. Gevert, S. Wang and R. U. Broze: Heat to a.c. conversion using thermoelectric generators and tunnel diodes 100
J. R. Dale and J. C. Brice: Etch pits in germanium and their relation to hardness . . 105
G. Strull and W. M. Kaufmann: An astable four-region multivibrator 110
J. M. Goldey, I. M. Mackintosh and I. M. Ross: Turn-off gain in p–n–p–n triodes . . 119
W. Rindner: A negative resistance photoelectric diode 123
T. E. Hartman: Transient photovoltaic response of diffused-junction silicon photodiodes . 127
D. F. Allison, R. H. Beeson and R. M. Shultz: kMc/s planar transistors in microwatt logic circuitry 134
W. Bardsley, J. M. Callan, H. A. Chedzey and D. T. J. Hurle: Constitutional supercooling during crystal growth from stirred melts—II 142

Note:
J. J. Moll and N. I. Meyer: Secondary multiplication in silicon 155

NUMBERS 3/4

H. L. Henneke: Indium antimonide transistors 159
Robert A. Pucel: Theory of the Esaki diode frequency converter 167
G. Dermit: Tunnel-diode power 208
David P. Kennedy and Philip C. Murley: Theoretical current gain of a cylindrical mesa transistor 215
T. C. Taylor and C. J. Bardsley: The selective melting of germanium by thermal radiation . 226
J. C. Danko, G. R. Kilp and H. M. Ferrari: An investigation of $Li_xNi_{1-x}O$ as a mixed-valence thermoelectric material 233
K. Landecker and A. W. Findlay: Study of the fast transient behaviour of Peltier junctions . 239
W. G. Pfann: Improvement of semiconducting devices by elastic strain 261
Charles N. Rollinger and J. Edward Sunderland: The performance of a convectively cooled thermoelement used for power generation 268

NUMBERS 3/4—*continued*

J. LINDMAYER and C. WRIGLEY: A simplified approach to transistor admittances . . . 278
J. SHAO and G. T. WRIGHT: Characteristics of the space-charge-limited dielectric diode at very high frequencies . . . 291
S. AMER: Influence of wafer thickness and carrier recombination on the cutoff frequency of alloy-junction transistors . . . 304

Notes:
N. HEMMAT and A. L. MACDONALD: Nomograph technique for doping determination in germanium and silicon crystal growing . . . 309
MILTON GREEN: Electrode geometries for which the transverse magnetoresistance is equivalent to that of a Corbino disk . . . 314
L. M. LAMBERT: Growing heavily compensated germanium crystals of known impurity concentrations . . . 316
D. T. J. HURLE, O. JONES and J. B. MULLIN: Growth of semiconducting compounds from nonstoichiometric melts . . . 317
R. N. HALL: Current gain in metal-insulator tunnel triodes . . . 320

Book Review:
Progress in Very High Pressure Research, by F. P. BUNDY, W. R. HIBBARD, JR. and H. M. STRONG (Reviewed by DOUGLAS WARSCHAUER) . . . 323
Encyclopaedic Dictionary of Physics, Vol. 1, Edited by J. THEWLIS (Reviewed by L. PINCHERLE) 324
British Miniature Electronic Components and Assemblies Data Annual 1961–62, Edited by G. W. A. DUMMER and J. M. ROBERTSON (Reviewed by ALAN F. GIBSON) . . . 325
Electronics: a Bibliographical Guide, by C. K. MOORE and K. J. SPENCER (Reviewed by ALAN F. GIBSON) . . . 325

Erratum:
M. GREEN, I. N. GREENBERG, J. BANZHAF, A. SASS, H. S. WHITE and R. V. MCKNIGHT: Electrical characteristics of a reduced-area, rhodium-plated, silicon end contact . . . 326

AUTHOR INDEX

(B.R. indicates a Book Review, E. an Erratum)

AHLBURG H., 11
ALLISON D. F., 134
AMER S., 304
BANZHAF J., 1, 326 (E.)
BARDSLEY W., 142, 226
BEESON R. H., 134
BRAUN I., 79
BRICE J. C., 105
BROZE R. U., 100
BUTTER C. D., 24
CALLAN J. M., 142
CHEDZEY H. A., 142
DALE J. R., 105
DANKO J. C., 233
DERMIT G., 208
DUNLAP W. C., 76 (B.R.)
EDWARDS D. F., 24
ERIKSEN W. T., 11
FA C., 18
FARRARI H. M., 233
FINDLAY A. W., 239
GEVERT F. G., 100
GIBSON A. F., 325 (B.R.s)
GOLDEY J. M., 119
GOODMAN G. H. L., 72
GREEN J. J., 77 (B.R.)
GREEN M., 1, 314, 326 (E.)
GREENBERG I. N., 1, 326 (E.)
HALL R. N., 320
HARTMAN T. E., 127
HEMMAT N., 309
HENNEKE H. L., 159
HEYWANG W., 51
HURLE D. T. J., 37, 142, 317
JONES O., 317
KAUFMAN W. M., 110
KENNEDY D. P., 215
KILP G. R., 233
KO W.-H., 59
LAMBERT L. H., 316
LANDECKER K., 239
LEHOVEC K., 45
LINDMEYER J., 278
LLOYD P., 74
MACDONALD A. L., 309
MACKINTOSH I. M., 119
MCGLAUCHLIN L. D., 24
MCKNIGHT R. V., 1, 326 (E.)
MEYER N. I., 155
MOLL J. J., 155
MULLIN J. B., 317
MURLEY P. C., 215
PARROTT J. E., 91
PENN A. W., 91
PFANN W. G., 261
PINCHERLE L., 324 (B.R.)
PUCEL R. A., 167
REDINGTON R. W., 76 (B.R.)
RINDNER W., 123
ROLLINGER C. N., 268
ROSS I. M., 119
SASS A., 1, 326 (E.)
SATO K., 29
SHAO J., 291
SHULTZ R. M., 134
SLOBODSKOY A., 45
STRULL G., 110
SUNDERLAND J. E., 268
TAYLOR T. C., 70, 226
TOMONO M., 29
WANG E. Y., 79
WANG S., 100
WARSCHAUER D., 323 (B.R.)
WHITE H. S., 1, 326 (E.)
WRIGHT G. T., 291
WRIGLEY C., 278
YUAN F. L., 70
ZULEEG R., 18

BACK NUMBERS OF PREVIOUS VOLUMES AVAILABLE

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 1–10. Printed in Great Britain

ELECTRICAL CHARACTERISTICS OF A REDUCED-AREA, RHODIUM-PLATED, SILICON END CONTACT

M. GREEN*, I. N. GREENBERG, J. BANZHAF, A. SASS, H. S. WHITE and R. V. McKNIGHT

U.S. Army Signal Research and Development Laboratory, Fort Monmouth, New Jersey

(*Received* 25 *October* 1960; *in revised form* 31 *January* 1961)

Abstract—The apparent linear (ohmic) characteristics, exhibited by a high-purity, p-type, single-crystal specimen of silicon of uniform resistivity (ca. 2200 Ω-cm at room temperature), having rhodium-plated end contacts, were changed to rectifying characteristics by formation of a reduced-area contact at one end (the reduced area was attained by grinding down the original contact). The rectification characteristics improved with decreasing temperature, provided that the temperature is not too low. Voltage-profile and resistivity measurements were made at various temperatures ranging from 116°C to −196°C. These curves show that the maximum of the injection fraction, $(\Delta n)/p_0$, and of the resistivity both occur at approximately 90°C. The values of these maxima are respectively 5·5 per cent (total current 10 mA, current density at small-area contact 125 mA/cm²) and 2980 Ω-cm. The low-temperature voltage contours indicate that the plated contacts, with both small and large area, become good barriers below −68°C. The asymmetry between the large- and small-area contact is much greater at −196°C than at −68°C. At room temperature, the transient behavior manifests the characteristic logarithmic injection curve. Solder and rhodium-plated contacts and transient-current studies indicate that a small amount of injection occurs also at the large-area contact.

Résumé—Les caractéristiques linéaires apparentes exhibées par un spécimen monocristallin de silicium de type p et ayant des contacts d'extrêmité plaqués et une résistance uniforme (2200 Ω-cm à température ambiante) ont ete changées en caractéristiques de redressement par la formation d'une surface de contact réduite à une extrêmité (la surface a été réduite en meulant le contact initial). Les caractéristiques de redressement se sont ameliorées en réduisant la température, si celle-ci n'est toutefois pas très basse. Des courbes de tension et des mesures de résistance ont été prises à diverses températures dans la gamme de 116°C à −196°C. Ces courbes démontrent que le maximum de la fraction d'injection $(\Delta n)/p_0$, et de la résistivité se produisent toutes deux aux environs de 90°C. Les valeurs de ces maxima sont respectivement 5,5 pour cent (courant total 10 mA, densité de courant à petite surface de contact, 125 mA/cm²) et 2980 Ω-cm. Les contours de tension à basse température indiquent que les contacts plaqués, ayant de petites et grandes surfaces, deviennent de bonnes barrières au-dessous de −68°C. L'asymétrie entre la grande et la petite surface est beaucoup plus grande à −196°C qu'à −68°C. A temperature ambiante, la façon transitoire de comportement dénote une courbe caractéristique d'injection logarithmique. Les contacts soudés et plaqués au rhodium et les études du courant transitoire indiquent qu'une petite quantité d'injection se produit aussi au contact à grande surface.

Zusammenfassung—Ein Silizium-Einkristall hohen Reinheitsgrades vom p-Type mit gleichförmigem spezifischem Widerstand (etwa 2000 Ω-cm bei Zimmertemperatur), der Rhodium-plattierte Endkontakte besass, zeigte eine scheinbar lineare (ohmsche) Kennlinie. Durch Herstellung einer kleinen Kontaktfläche (wir erzielten diese durch Abschleifen des ursprünglichen Kontaktes) entstand eine Gleichrichterkennlinie, die mit abnehmender Temperatur besser wurde, vorausgesetzt, dass die Temperatur nicht zu niedrig war. Die Spannungsprofile und der spezifische Widerstand wurden in einem Temperaturbereich von 116 bis −196°C bestimmt. Die Kurven zeigen, dass das Maximum des Injektionsbruches $(\Delta n)/p_0$ sowie das des spezifischen Widerstandes bei etwa 90° liegen. Die Maximalwerte betragen 5,5 Prozent (Gesamtstrom 10 mA, Stromdichte am reduzierten Kontakt 125 mA/cm²) und 2980 Ω-cm. Die Spannungsprofile bei niedrigen Temperaturen deuten an, dass plattierte Kontakte, sowohl mit grosser als mit kleiner Oberfläche, unter-

* Present address: Burroughs Research Center, Paoli, Pa.

halb von −68° gute Sperren bilden. Der Unterschied zwischen grossen und kleinen Kontaktflächen ist bei −196° viel ausgesprochener als bei −68°. Bei Zimmertemperatur ergibt sich für den Durchgang die charakteristische logarithmische Injektionskurve. Untersuchungen an gelöteten und Rhodium-plattierten Kontakten und Beobachtungen des Durchgangsstromes zeigen, dass auch bei grossen Kontaktflächen ein geringer Grad von Injektion besteht.

INTRODUCTION

IT IS well known that the concentration of hole–electron pairs in a semiconductor may increase or decrease[1] from the equilibrium value by various processes. One of the most common methods of increase is injection of minority carriers at a p–n junction or at a metal contact with the semiconductor surface. The theories for these phenomena have been developed by various workers and are discussed by SHOCKLEY[2]. DUNLAP[3] mentions that contact rectifiers are of two types: (a) those in which differences in contact area are responsible for rectification, and (b) those in which differences in contact properties are responsible. This paper is concerned mainly with type (a) rectification, and the experimental investigation deals primarily with a study of the variation with temperature of small area (ca. 0·08 cm^2 formed at one end of the sample previously rhodium-plated by grinding off part of the silicon and plating in the shape of a parallelepiped 1·1 cm long × 0·07 cm wide × 0·04 cm deep) versus large-area (ca. 1 cm^2) injection-and-blocking effects of contacts of rhodium plated to high-purity p-type, single-crystal silicon, in the form of a cylindrical rod 1·1 cm in diameter by 2·0 cm long, of uniform resistivity (ca. 2200 Ω-cm at room temperature). Interpretations of the results are given.

RESISTANCE MEASUREMENTS

The sample, described above, was prepared by first grinding the ends so they were parallel to each other and had uniform, flat surfaces. The ends were then abraded with a fine abrasive paper or cloth, like Norton 180–C, so that the electrodeposited contact would adhere better. The sample was plated cathodically versus a platinum anode by just allowing the end surface to touch the surface of the solution and then withdrawing it slightly so that a meniscus was formed in contact with the surface to be plated. The bath was dilute acid (0·02 ml H_2SO_4/l.) rhodium-plating solution containing 2 g/l. of rhodium. A plated contact of about 5 μ thickness was formed at a current density of 10 mA/cm^2 in about 30 min. The contact was examined, visually, with a magnifying lens for obvious flaws such as pin-holes, non-uniformities, flaking, etc. The surface resistance of the plate was also probed along the contact to make sure there were no high-resistance spots or discontinuities in the plating which were missed in the visual examination. Both end contacts were of equal area (ca. 1 cm^2) at this point. The sample was scanned with a single-probe resistivity apparatus and was found to have uniform resistivity (ca. 2200 Ω-cm) at room temperature. A reduced-area contact was then formed by grinding down one contact as described above. This specimen was used for all measurements except those where the effect of a solder contact was studied.

Except as stated below, resistance measurements were taken with a Wheatstone bridge with an external battery supply. The bridge current was read with a milliammeter in series with the battery and the actual sample current was calculated from the known bridge arm resistances. Power to the bridge was keyed to prevent heating of the sample. The resistance bridge was too inaccurate for the high resistances measured at −70°C and −196°C. In this case the resistance was measured by taking the ratio of the voltage V across the specimen to the current I flowing through it. Since the resistance of the sample decreased in time when the current was flowing through it (because of joule heating), it became necessary to get the initial value of I, i.e. before the sample temperature had changed appreciably, so that a high speed recording microammeter had to be employed for this purpose.

Most probably the primary heating effect occurs at the small area contact. The concentrated temperature rise near this contact as a consequence of the large current density could partly explain the observations reported below. On the other hand, the conditions of the experiment were: the power to the sample was small (less than 0·1 cal/sec), the thermal conductivity of silicon was good (of the order of 1–2 W/cm-deg), and the

sample was in good thermal contact with the bath (a good heat sink). It would be expected that any temperature gradient in the sample due to the primary heating at the reduced-area contact would be cut down rapidly and considerably. The temperature measurements indicated that the sample temperature was quite uniform and the temperature increase due to the I^2Rt (t is time) effect occurred sufficiently slowly to permit measurement of the sample resistance, at a particular known temperature, by the method stated above.

The results of the measurements are shown in Figs. 1 and 2. A rectification characteristic is observable in all cases. At −196°C the high-resistance branch of the curve (polarity of reduced-area contact positive) decreases with increasing current (see Fig. 2) in contrast to the curves at the other temperature where it increases. The cause for this will be clearer after the data in the section on voltage profile and resistivity have been discussed.

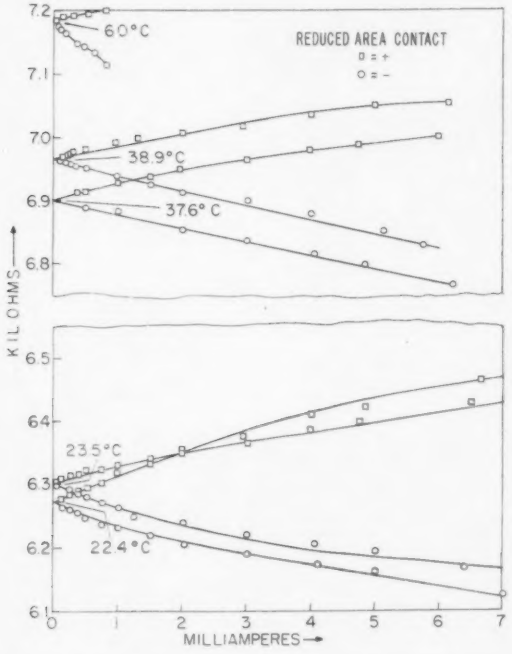


FIG. 1. Rectification of rhodium-plated, reduced-area end contact on p-type silicon, at room temperature and above.

VOLTAGE PROFILE AND RESISTIVITY MEASUREMENTS

Voltage profiles were taken at various temperatures, Figs. 3–5, using a single-probe resistivity apparatus. The current across the crystal was kept constant for each run. The voltage increments were measured by using a bucking voltage to obtain the differential voltage directly, so as to obviate the inaccuracy introduced by taking the small difference between two large numbers. The ambient temperature was kept as constant as conditions would allow (variations of a few degrees were observed). Deviations from the average temperature were small for each particular contour determination. No thermal gradients were observed in the sample. The measurements were made after temperature equilibrium had been reached.

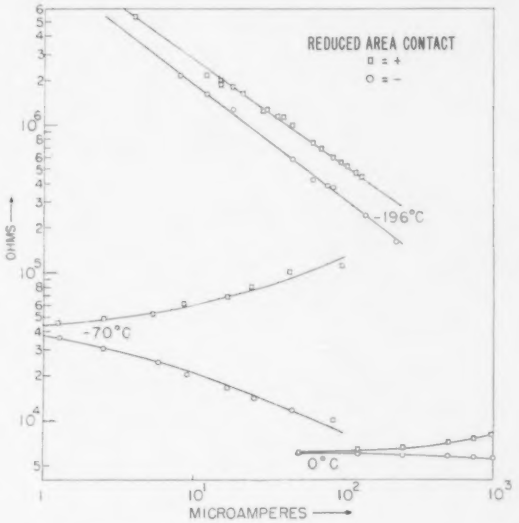


FIG. 2. Rectification of rhodium-plated, reduced-area end contact on p-type silicon, below room temperature.

The large field due to the reduced-area contact extends approximately 2 mm from the small-area end contact at room temperature (26°C), Fig. 4. The reduced-area contact constrains the current so that in its immediate vicinity the current density is high. Since the electric field in the bulk material varies inversely as the conductivity for a given

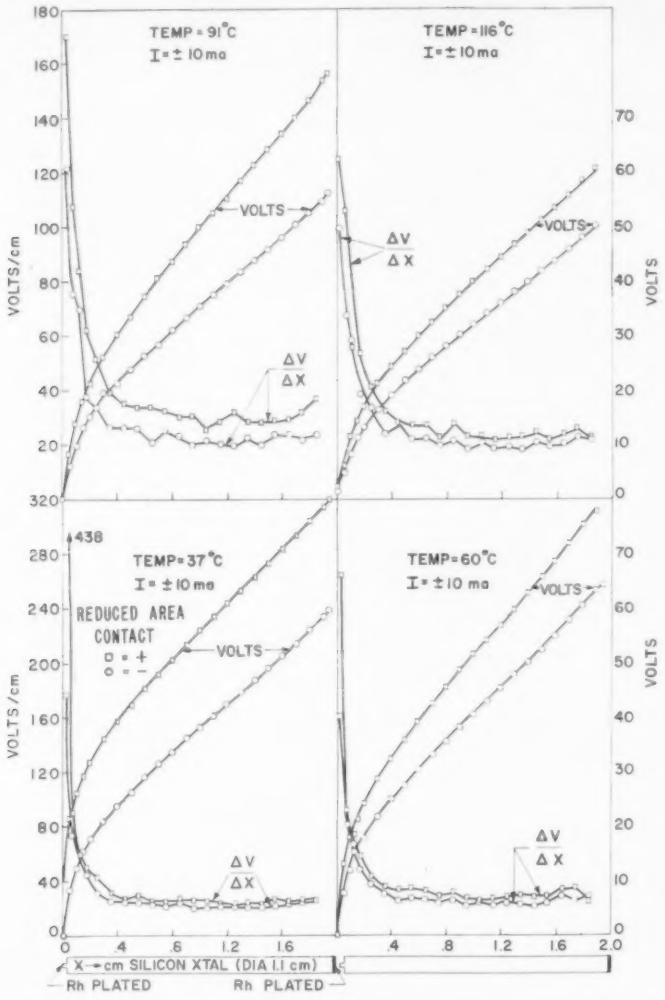


FIG. 3. Voltage-profile curves, above room temperature, showing rectification and injection of rhodium-plated, reduced-area end contact on p-type silicon.

current density, and the conductivity increases with injection, differences in field strength for the current flowing in opposite directions gives a measurement of the difference in injection* at

* After BRAY[1]. Equation (8) applied for the conditions cited in this text:

$$\frac{\Delta\sigma}{\sigma} = \frac{-\Delta E}{E} = (\gamma_{c_1}-\gamma_0)\frac{b+1}{b} - (\gamma_{c_2}-\gamma_0)\frac{b+1}{b} = (\gamma_{c_1}-\gamma_{c_2})\frac{b+1}{b}; \text{ for } J = \text{const.}$$

where $\Delta\sigma$ is the change in conductivity due to differences in injection for the two directions of current flow, $-\Delta E$ is the corresponding change in the field strength, σ is the conductivity, E is the field strength in the vicinity of the contact, γ_c is the ratio of minority-carrier current to the total current at the contact, γ_0 is the same ratio in the interior and is a constant, and b is the electron-to-hole mobility ratio. J is the current density.

the two different end contacts, provided that the transit time of minority carriers is short compared with carrier lifetime. Thus, for Figs. 3–5 the displacements of the field profiles at a distance from

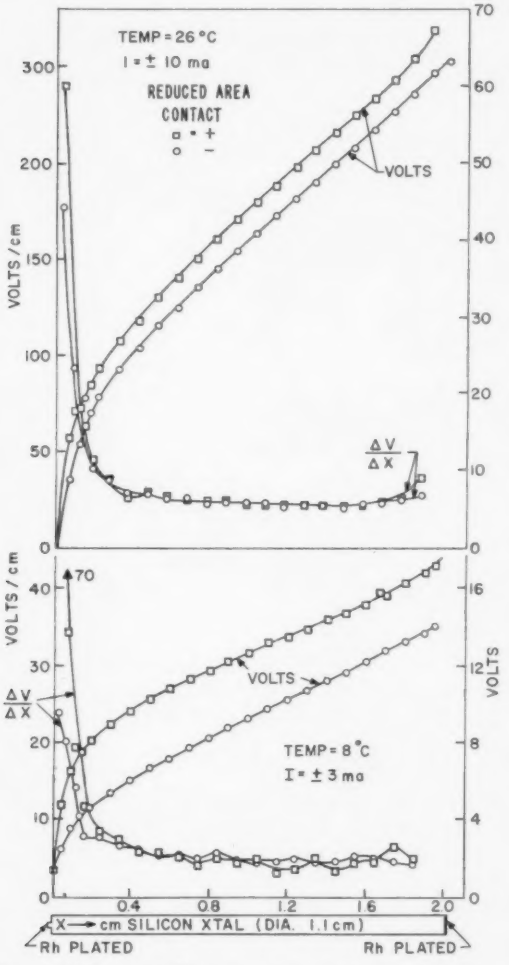


FIG. 4. Voltage-profile curves, near room temperature, showing rectification and injection of rhodium-plated, reduced-area end contact on *p*-type silicon.

the junction are fair measures of the injection difference between the two end current contacts. For comparison, near room temperature, using a similar silicon sample with uniform cross section and with a longer lifetime (ca. 1100 μsec), with one contact ultrasonically soldered (60–40 solder),

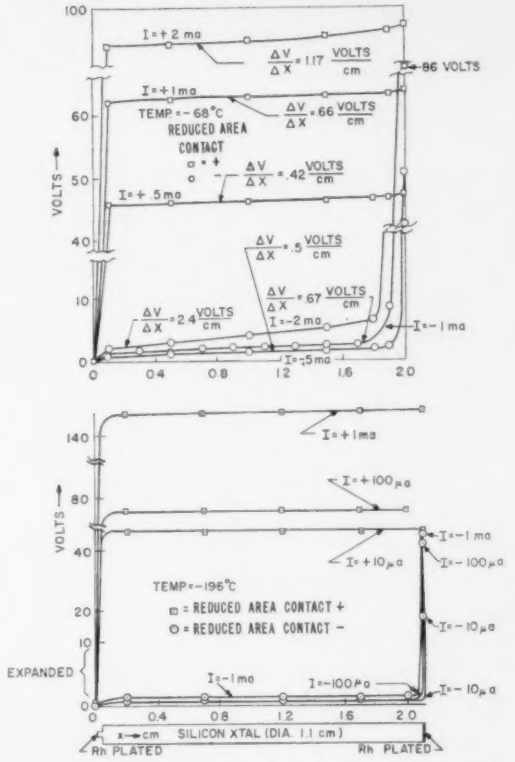

I = + 2 ma
ΔV/ΔX = 1.17 VOLTS/cm
86 VOLTS
I = + 1 ma
TEMP = − 68 °C
REDUCED AREA
CONTACT
□ = +
○ = −
ΔV/ΔX = .66 VOLTS/cm
ΔV/ΔX = .42 VOLTS/cm
I = + .5 ma
ΔV/ΔX = .5 VOLTS/cm
ΔV/ΔX = .67 VOLTS/cm
ΔV/ΔX = 2.4 VOLTS/cm
I = − 2 ma
I = − 1 ma
I = − .5 ma
VOLTS
I = + 1 ma
I = + 100 μa
I = − 1 ma
I = + 10 μa
I = − 100 μa
TEMP = − 196 °C
□ = REDUCED AREA CONTACT +
○ = REDUCED AREA CONTACT −
I = − 10 μa
EXPANDED
I = − 1 ma
I = − 100 μa
I = − 10 μa
x → cm SILICON XTAL (DIA. 1.1 cm)
Rh PLATED
Rh PLATED

FIG. 5. Voltage-profile curves at −68°C and −196°C demonstrating barrier at rhodium-to-silicon for both large- and small-area contacts.

and the other, of equal area but rhodium plated, the voltage and field profile curves, Fig. 6, show an obviously large amount of injection at the solder contact. Hence it is seen that from room temperature (26°C) to 91°C there is a gradual increase in injection. This is more clearly demonstrated by Fig. 7, where the curve shows the fractional change in conductivity (determined from the differences in the slopes of the voltage-profile curves) and also the ratio of pairs injected to hole concentration for the reverse direction. At the ambient temperatures of liquid nitrogen (−196°C) and dry ice (−68°C) the voltage-profile curves of the uniform part of the crystal are too inaccurate for quantitative measurement of the injection, although a comparison of the slopes of the curves for equal currents at −68°C does indicate some injection. There is undoubtedly

considerable injection at either contact when appropriately biased. This is discussed later.

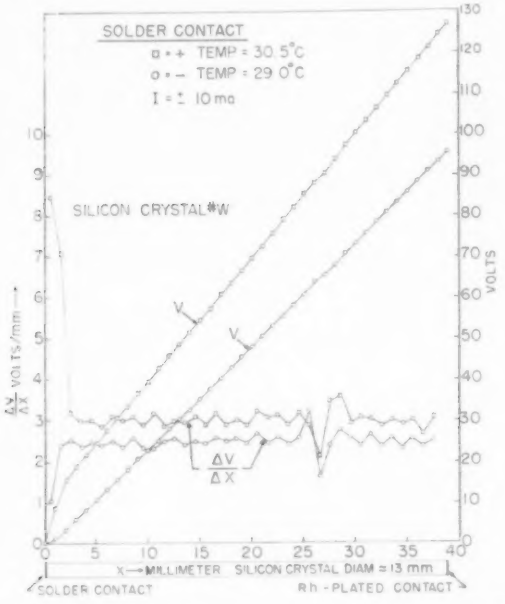


FIG. 6. Voltage-profile curves of crystal *W*, showing rectification and injection due to difference in contact properties, one contact 60–40 solder applied ultrasonically, the other, of equal area, rhodium plated. The effect of the temperature difference on the conductivity is 1·5 per cent while the over-all change in conductivity of the bulk material is 15 per cent corresponding to a net injection $\Delta n/p_0$ of about 3 per cent.

Some voltage-profile curves were also made with a lower current of 100 μA in the temperature range $-5{\cdot}5$°C to -70°C, Fig. 8. There was a slight amount of instability in the d.c. measurements with this low current and at these temperatures. The ensemble of curves, however, is indicative of the direction in which the end effects and the slope of the uniform portion of the curves progress with decreasing temperature. The upward bending of the voltage-profile curves, in going from the injecting contact to the collecting contact, is indicative of a situation in which the minority-carrier lifetime is shorter than the transit time. This has yet to be proved.

SAMPLE RESISTIVITY VERSUS TEMPERATURE

Sample resistivities were determined from the

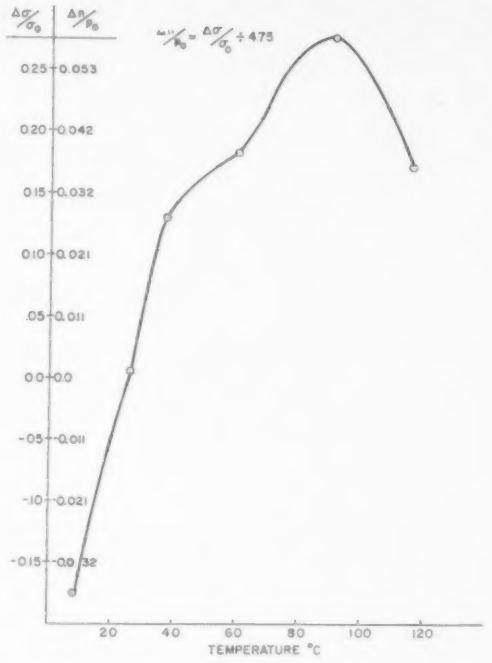


FIG. 7. Fractional change in conductivity $\Delta\sigma/\sigma_0$ and injection fraction $\Delta n/p_0$ vs. temperature, determined from the differences in the slopes of the voltage-profile curves.

voltage-profile curves and, at the two lowest temperatures, also by a two-probe method. A plot of the data is shown in Fig. 9. The data are in good agreement with the results of PUTLEY and MITCHELL's[4] specimen ZG133. The slope of the curve in the exhaustion region gives a temperature dependence of $T^{-2{\cdot}67}$. For comparison, PUTLEY and MITCHELL[4] give a Hall mobility variation in this range of $T^{-2{\cdot}7}$. Their curves show that the Hall coefficient is substantially constant in this temperature range.

TEMPERATURE DEPENDENCE OF THE VOLTAGE OF THE BLOCKING LAYER

Voltage versus temperature measurements were made on the specimen with 1 mA of current and small-area contact negative (injecting polarity). The reproducibility of the measurements was only fair. A curve* approximating the average results is

* A somewhat similar curve is found for photovoltage vs. temperature[5].

Table 1. *Variation of sample resistance with current**

Temp. (°C)	Current (μA)	Voltage (V)		Resistance (kΩ)	
		Forward	Reverse	Forward	Reverse
−68	2000	83	95	41·5	47·5
	1000	50·5	64	50·5	64
	500	42·0	47	84	94
−196	1000	45	148	45	148
	100	42	75	420	750
	10	20	47	2000	4700

* Data are same as in Fig. 5.

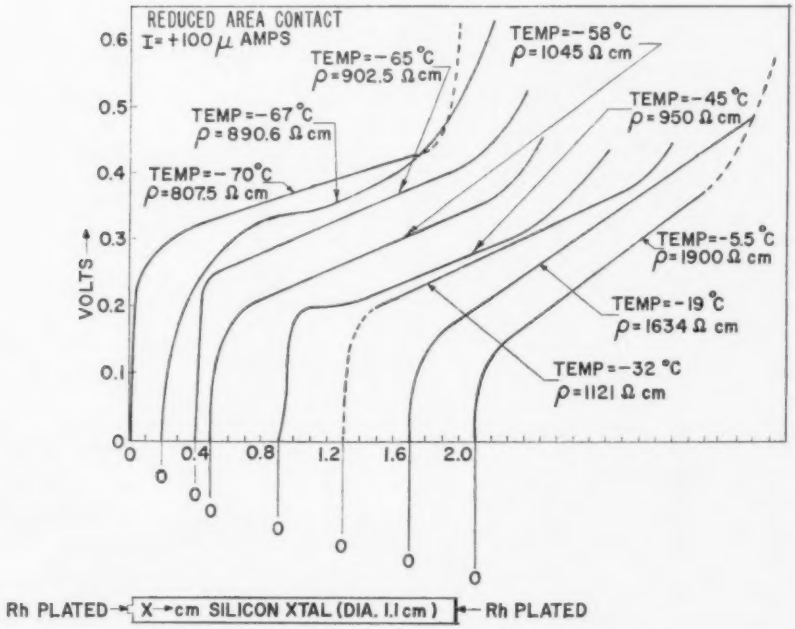


FIG. 8. Variation of voltage-profile curves (100 μA current) with temperature.

shown in Fig. 10. The steep rise in the curve, with decreasing temperature, occurring in the vicinity of −50°C corresponds to the temperature at which, theoretically, almost all of the minority carriers should be frozen out, Fig. 11. According to the theory* of the p–n junction, the reverse current through a junction is practically all carried by minority carriers. One might expect

* Ref. 2, Sections 4.2, 4.3 and 12.5.

this to apply also to the case of a p-semiconductor to a *metal* junction. If this is the case, then at temperatures below minority-carrier freeze-out, the blocking layer should be depleted of free carriers for reverse conduction*.

To account for the low-level currents† observed in Figs. 2 and 8, in the absence of minority carriers, another mechanism of conduction must be postulated. Possibilities that have appeared in the literature are impurity-band conduction,[6] which in this case is not very likely, and internal-field emission (tunneling)[7]. It is beyond the scope of this paper to go into these in great detail; however, a few rather speculative remarks will be ventured with respect to internal-field emission as a possible explanation for the low-temperature resistivities observed.

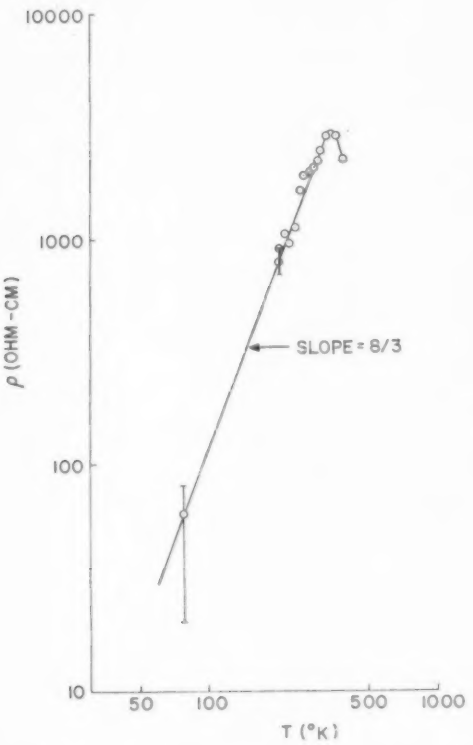


FIG. 9. Resistivity vs. temperature. The slope of the curve in the exhaustion region corresponds to a temperature dependence of $T^{-2.67}$.

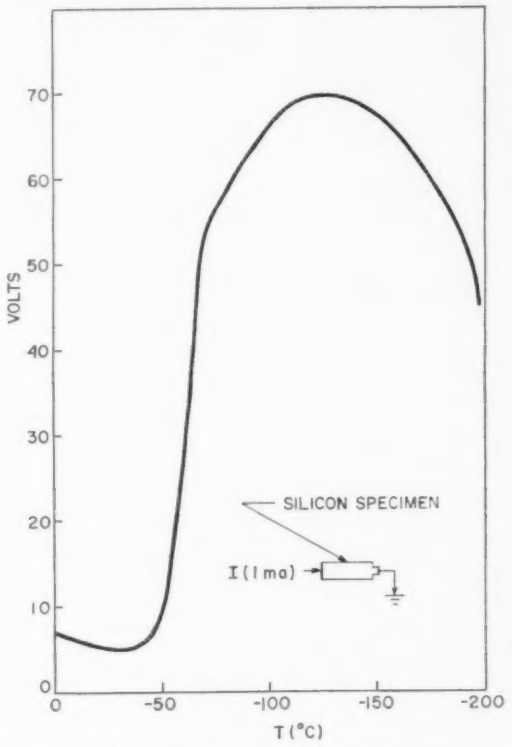


FIG. 10. Voltage across specimen vs. temperature at 1 mA current, reduced-area contact negative.

One aspect of this concept might be viewed as a type of reverse-bias tunneling[7] via acceptors, or just direct tunneling. A physical picture of the case of tunneling via acceptors would be somewhat as follows.

The reverse bias that causes the depletion layer leaves it without holes, and, since the electrons are frozen out, also without electrons. In this "layer" there will be considerable space charge. On the

* For example, at $T = 225°K\ (-48°C)$ the bulk electron concentration of our silicon specimen is approximately 10 per cm^3, Fig. 11, and the hole concentration is $p = (N_A - N_D) = 6 \times 10^{12}$ per cm^3. Assuming a very generous diffusion velocity, $V_D = \sqrt{(D/\tau)}$, of the order of 10^5 cm/sec for electrons, the saturation-current density, $I_s = en\sqrt{(D/\tau)}$, would be of the order of 10^{-13} A/cm^2.

† Once the current reaches a high enough value, so that minority-carrier transit time is less than the lifetime, injection at the contact opposite the blocking one can account for a considerable part of the conductivity.

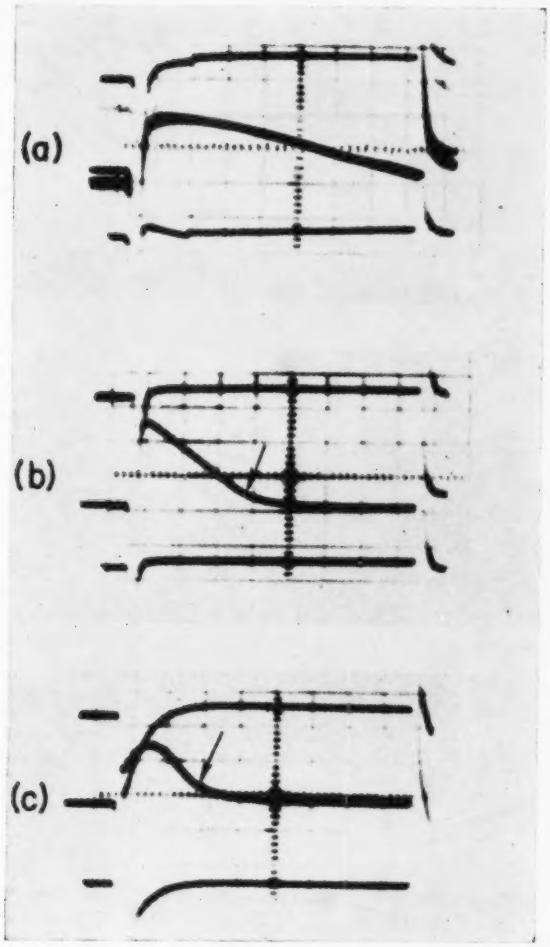


FIG. 12. Transient-current studies of injection in p-type silicon rhodium-plated, reduced-area end contact. Arrival time of minority carriers at opposite contact is shown by kink (arrow) in the center traces of (b) and (c). For (a), (b) and (c); top curves, small-area contact +, center curves, small area contact −, and bottom curves, commercial resistor. Scales: horizontal = 10 μsec/major div., vertical = (a) 50 mV/major div., (b) 2 V/major div., (c) 500 mV/major div. Applied pulse voltage to bridge = (a) 50 V, (b) 135 V and (c) 210 V.

[facing p. 8

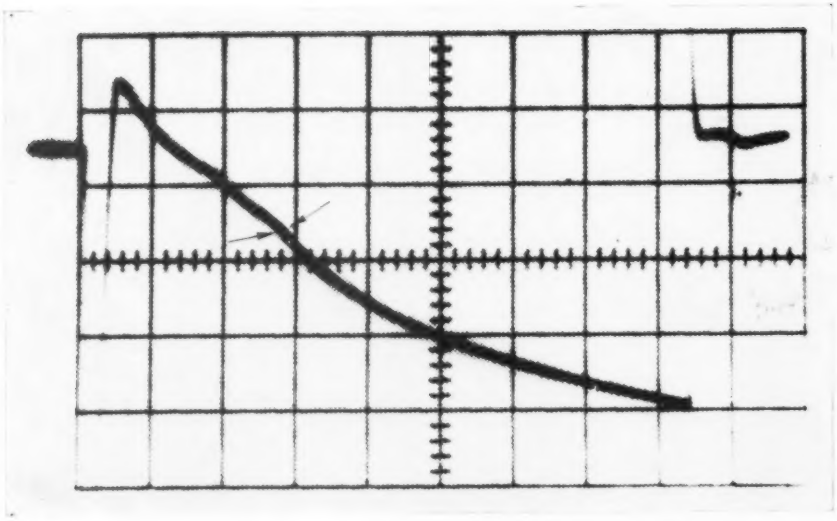


FIG. 13. Transient-current studies of reverse-bias inection in p-type silicon at the broad-area contact. First kink due to pulser, second kink (arrows) shows arrival time of minority carriers at reduced-area contact. Voltage across sample = 200 V. Scales: horizontal = 10 μsec/major div.; vertical = 100 mV/major div.

semiconductor "edge" this would consist mainly of acceptors. The probability of ionizing an acceptor completely, or of transfer of its electron

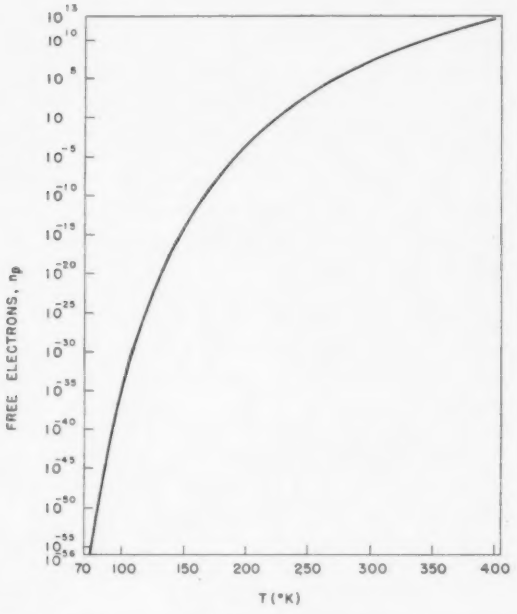


FIG. 11. Log n_p (free-electron concentration) vs. temperature. Calculated from[5] $n_p = n_i^2 = 9{\cdot}6 \times 10^{32} T^3 \times \exp(-1{\cdot}206/kT)$.

to an unfilled donor state, will, of course, depend on the energy difference between* the two states, and on the density of acceptors; and for transitions to a donor state, on the density of ionized donors. The transition probability will also depend very strongly on the value of the local electric field[8] E at the acceptor. Once the transitions start, conduction proceeds by transfer of electrons from the p-side to the metal via high electric-field ionization of acceptors alone, or via acceptor ionization (very unlikely here, but a possibility) via donor capture and re-emission. The electrons in the valence band then refill the vacated acceptors with the formation of holes. The process then becomes self-sustaining.

A field-emission process that increases with electric field, or a combination of this process and of injection at the opposite contact would readily explain the decrease in resistance with current (both polarities) of the curves at −196°C, Fig. 2. For the range of currents shown, the curve for −70°C exhibits an apparent "normal" type of behavior. However, at higher currents, Fig. 5 and Table 1, the effects are similar to those observed at −196°C. A maximum occurs in the curve, Fig. 10, in the vicinity of −125°C. The interpretation is that this is the temperature at which there is a minimum number of electrons in the junction when 1 mA of current is flowing for a current contact area of 0·08 cm². This is a complex resultant of the effects of electron freeze-out in the bulk material, electron injection into the bulk material dependent upon lifetime, electron transport through the bulk material dependent upon transit time, and the "internal field emission" at the junction to the material.

INJECTION OBSERVED THROUGH TRANSIENT-CURRENT STUDIES

The transient currents resulting from application of a step voltage were examined at room temperature with a simple bridge circuit similar to that used by BRAY[1] in his investigation of minority-carrier extraction.

The effects on the transient current of pulse-voltage amplitude and reduced-area contact polarity are shown in Fig. 12 where the horizontal axis is time, in microseconds per major division, and the vertical axis corresponds to increase in resistance. The top, center and bottom curve of each oscillogram correspond to transients in which, (1) small-area contact is positive, (2) small area contact is negative, and (3) a 2-W, composition, IRC commercial resistor has been substituted for the crystal. The pulse voltage applied to the bridge, only about half of which appears across the silicon crystal, is given for Fig. 12(a–e) as, 50, 135 and 210 V, respectively. It is observed in each case that the top and bottom curves—corresponding, respectively, to reduced-area contact positive and to the commercial resistor—are nearly horizontal, indicating that there is little or no

* KELDYSH[8] gives, for the field and energy-dependent factor in the probability of tunneling, $\exp(-\alpha\epsilon^{3/2}/E)$, where α is a constant depending on the square root of the effective mass of the carrier and ϵ is the energy of the gap.

injection* of carriers for this small-area contact polarity. The center curves, on the other hand, exhibit negative slopes indicating decreasing resistance. In Fig. 12(b) and (c) the abrupt changes in the slope of these curves, or kinks, indicate the time at which the injected minority carriers have arrived at the opposite contact.

The above results indicate (1) that there is injection at the rhodium-to-silicon contact for the reduced-area contact and also for the large-area contact, (2) that the effect of the much greater current density and correspondingly greater voltage gradient at the small-area contact produces an enormously greater total injection there. Similar transient experiments have been performed on the piece of high-purity silicon of Fig. 6 with contacts of equal area, but with one plated with rhodium and the other soldered ultrasonically with 60–40 solder. Soldering forms an alloy junction at the contact. The transient curves are quite similar, except that the injection at the large-area rhodium contact is more easily observed in this specimen than in the other having the small-area contact†, the kink in the curve indicating the arrival time for carriers injected at the rhodium contact being more pronounced.

CONCLUSIONS

(1) Rectification effects observed as a result of the difference in the area of the two current contacts of rhodium plated to high-purity p-type silicon are explained qualitatively. The implication is that the injection and extraction are non-linear functions of the current density (or electric field) at the junction. The concentrated temperature rise near the small-area contact may be partly responsible for the effects observed, but under the conditions of the experiment, it does not appear very likely.

(2) The rectification characteristic is highly temperature dependent.

(3) Conduction mechanisms other than the normal hole conduction in the valence band or electron conduction in the conduction band must be considered as a means of low-level conduction through the blocking layer at temperatures in the exhaustion range and below minority-carrier freeze-out.

(4) Transient-injection studies demonstrate that injection occurs at both the small- and large-area rhodium-to-silicon contacts when the proper polarity is used. The evidence of injection at the large-area contact is supported by the low-temperature data and the data shown in Fig. 6.

* Actually when the vertical scales of these oscillograms are greatly enlarged, a slight amount of injection is observable which is attributable to large-area rhodium contact (small-area contact). The transit kink is also observed, Fig. 13, horizontal scale 10 μsec/major div., vertical scale 100 mV/major div. Although not shown in Figs. 12 or 13, evidence of the rapid formation of the blocking layer (reduced-area contact positive) is also observed, as demonstrated in the following manner. The bridge is first balanced with d.c. at a very low current level. Also, the bridge is then pulsed with 50 V or more. A large unbalance of the bridge is observed indicating a substantial change in specimen resistance. Unfortunately, the leading edge of the pulse used, while quite steep, is not sufficiently clean and, when coupled with a slight amount of unbalanced distributed capacitance, does not permit a quantitative temporal study of the blocking layer formation. When the specimen is replaced with a commercial resistor, the balance of the bridge remained the same for both d.c. and pulsed voltages.

† This is because the reverse-biased solder-contact junction acts as a better indicator of the arrival of the injected carriers than the small-area contact, i.e. there is a larger percentage change in resistance at the contact.

As shown in Fig. 6, the solder contact is the better-injecting contact. In the above discussion, however, a comparison is made between small-area contact and large-area contact injection at a fixed total current for rhodium plated to p-silicon.

REFERENCES

1. R. Bray, *Phys. Rev.* **100,** 1047 (1955); P. C. Banbury, *Proc. Phys. Soc. Lond.* **B66,** 50 (1953).
2. W. Shockley, *Electrons and Holes in Semiconductors* Chaps. 4 and 12. Van Nostrand, New York (1951).
3. W. C. Dunlap, Jr., *An Introduction to Semiconductors* p. 130. Wiley, New York (1957).
4. E. H. Putley and W. H. Mitchell, *Proc. Phys. Soc. Lond.* **72,** 195 (1958).
5. M. Green, I. N. Greenberg, J. Bonzhaf, A. Suss, H. S. White and R. V. McKnight. To be published.
6. C. S. Hung and J. R. Gliessman, *Phys. Rev.* **96,** 1226 (1954); C. Herring, *Ibid.* **96,** 1163 (1954) (see footnote 11, p. 1164).
7. L. Esaki, *Phys. Rev.* **109,** 603 (1958).
8. F. V. Keldysh, *J. Exp. theor. Phys.* **33,** 994 (1957); **34,** 962 (1958). Translations: *Soviet Phys. JETP* **6,** 763 (1958) and **7,** 665 (1958), respectively.

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 11–17. Printed in Great Britain

AN ANALYSIS OF ELECTROLUMINESCENT BRIGHTNESS WAVES

W. T. ERIKSEN* and H. AHLBURG

Research Division, Raytheon Company, Waltham, Massachusetts

(*Received* 10 *August* 1959; *in revised form* 5 *January* 1961)

Abstract—Using the concept of impact ionization as the means of excitation of electroluminescence and making an assumption concerning the phase relationships between applied voltage and exciting electrons, an equation has been derived which gives the time dependence of the instantaneous brightness.

Solutions of this equation are compared to experimental brightness waves and similarities pointed out. The main result is that the existence of secondary peaks in the brightness waves is predicted without reference to trapping phenomena or delayed recombination.

Résumé—En employant la conception d'ionisation de choc comme moyen d'excitation d'électroluminescence et en assumant une hypothèse concernant les relations de phase entre la tension appliquée et les électrons excités, une équation a été dérivée indiquant l'éclat instantané en fonction du temps.

Les solutions de cette équation sont comparées aux ondes d'éclat expérimentales et les similarités sont indiquées. Le résultat principal de cette étude est de prédire l'existence de crêtes secondaires des ondes d'éclat sans se référer aux phénomènes de trappe ou de recombinaison retardée.

Zusammenfassung—Unter Verwendung des Stossionisationsmechanismus zur Erregung der Elektrolumineszenz und mittels einer Voraussetzung über die Phasenbeziehung zwischen Spannung und anregenden Elektronen wurde eine Gleichung abgeleitet, die die Zeitabhängigkeit der momentanen Lichtintensität angibt.

Die Lösungen dieser Gleichung werden mit experimentellen Intensitätswellen verglichen und auf Übereinstimmung wird hingewiesen. Das Hauptergebnis besteht darin, dass die Existenz von sekundären Spitzen der Intensitätswellen vorauszusagen ist, ohne auf Einfangsphänomene oder verzögerte Rekombination Bezug zu nehmen.

I. INTRODUCTION

ONE of the interesting aspects of electroluminescence is the time dependence of instantaneous brightness of a phosphor as a function of the exciting potential wave form. With sinusoidal excitation there are two main peaks in brightness per cycle. Many phosphors show secondary peaks which vary in amplitude and position with respect to the main peak. These variations are affected by the temperature, the frequency of the exciting potential, and the properties of the phosphor under investigation.[1–8] The secondary peaks have been attributed to trapping phenomena, or diffusion of electrons back into the space-charge region, which introduce delayed recombination.† The proposed theories do not allow a concise statement of the properties of brightness waves.

In this paper an assumption is made about the mechanism of excitation of electroluminescence from which the existence of secondary peaks follows naturally. An equation will be derived from which the instantaneous brightness as a function of time is calculated for sinusoidal excitation. The same equation yields an expression

* Now with Semiconductor Division, Raytheon Company, Newton, Mass.

† An attempt to explain secondary peaks without using some delay mechanism has been made by E. E. LOEBNER and I. J. HEGYI at the 1959 Meeting of the Electrochemical Society in Philadelphia.

which contains the frequency dependence of average brightness at low frequencies.

II. FORMULATION OF BRIGHTNESS EQUATION

Excitation of electroluminescence by a.c. fields in zinc sulfide phosphors is assumed to be the result of impact ionization of luminescent centers by free electrons which have been accelerated to optical energies by high fields in the host crystal.[9,10] Controversy arises in discussion of the origin of the free electrons and the location and reason for the high-field regions in the crystal. PIPER and WILLIAMS[9] postulate deep-lying donor levels which are ionized by the high field as a source of the exciting free electrons. TAYLOR[10] agrees in the main but prefers to consider the deep-lying donors as thermally ionized. On the other hand, ZALM[3] assumes the source of the free electrons in a different phase at the surface of the crystal. There is general agreement that the high-field regions are obtained across the potential barriers either at a surface in a Mott–Schottky exhaustion region or across p–n junctions in the bulk, or both.

The assumption made in this paper is that, to a first approximation, for sinusoidal voltage input, the portion I of the current which is due to electrons moving in the regions where excitation is occurring also varies sinusoidally with time, having the same frequency but not necessarily the same phase as the exciting potential.

The rate dN/dt at which electroluminescent centers are excited is dependent on this current I, the supply of free electrons, and the value F of the field in the region where excitation is occurring.

$$\frac{dN}{dt} = f(I, F) \tag{1}$$

The functional dependence is obtained[10] by making use of SEITZ' relationship from his analysis of multiplication and breakdown in crystals[11]

$$P = \exp(-b'/F) \tag{2}$$

which states the probability that an electron moving in field F will have sufficient energy E to ionize a luminescent center. b' contains primarily E and a collison frequency with phonons. The rate at which luminescent centers are excited will then be proportional to the product of this probability and the rate of electron supply

$$\frac{dN}{dt} = A'|I| \exp(-b'/F) \tag{3}$$

A' is a constant containing such things as the collision cross section of the luminescent centers and their concentration. We now make the substitution[9] that for an abrupt potential barrier the field in the space-charge region is proportional to $V^{1/2}$.

$$\frac{dN}{dt} = A'|I| \exp(-b/V^{1/2}|) \tag{4}$$

We have placed absolute-value signs around symbols for current and voltage because by definition the excitation term can be only positive or zero. A negative and complex dN/dt is excluded. The fact that experimental brightness waves show equivalent instantaneous brightness over half cycles regardless of whether the potential and current are positive or negative, also indicates that only the absolute values of current and voltage have to be considered here. We are, however, neglecting the unequal magnitude of brightness on alternate half cycles which is frequently observed and has been explained by DESTRIAU[12].

The excited luminescence centers produce photons while decaying to the ground state with a characteristic relaxation time τ. For simplicity we assume this decay to be monomolecular[13], but our equations can easily be modified to apply to other types of decay.

We then use the following simple reaction scheme:

$$\text{ground state} \xrightarrow{A'|I| \exp(-b/V^{1/2}|)} N_e$$

$$\text{excited centers} \xrightarrow{N_e/\tau} \text{ground state} + h\nu.$$

This leads to an equation for the net change in the number of excited centers with time

$$\frac{dN_e}{dt} + \frac{N_e}{\tau} = A'|I| \exp(-b/|V^{1/2}|) \tag{5}$$

Since each excited center emits a photon when returning to ground state, the instantaneous

brightness B is N_e/τ in photons/sec. If the sinusoidal time dependence of V and I are substituted into this equation the time dependence of N_e/τ can be calculated.

III. SOLUTION OF BRIGHTNESS EQUATION

In order to solve equation (5), we need the time dependence of I. When a time-varying potential is applied to an electroluminescent panel, a current flows. It is made up of several components, capacitive and resistive, all bearing certain time relationships to the applied potential. The electrons in which we are interested here form part of this current. We then have the problem of finding the time dependence of this portion I of the total current. A complete solution to the problem posed here is very difficult. To simplify the situation, we assume that I is proportional to the applied sinusoidal potential but with a certain phase shift (a behavior similar to that of the total current through the panel).

If the applied voltage is $V = V_0 \sin \omega t$, we therefore use $I = I_0 \sin(\omega t+\theta) = I_0 \cos(\omega t-\phi)$. $\phi = \pi/2-\theta$ is a phase angle relating the flow of exciting electrons to the applied potential. It is not immediately obvious how ϕ can be obtained by independent experiments. It will be considered as an experimental parameter.

Making these substitutions, equation (5) can be written

$$\frac{dN_e}{dt} + \frac{N_e}{\tau} = A' I_0 \times \times \exp[-b/(V_0^{1/2}|\sin^{1/2}\omega t|)]|\cos(\omega t-\phi_1| \tag{5'}$$

an equation which cannot be integrated in terms of elementary functions since the generation term is too complicated. To obtain a useful solution, we have therefore taken a simple power of V to approximate the exponential term.

$$\frac{dN_e}{dt} + \frac{N_e}{\tau} = AI_0 V_0^m |\sin^m \omega t \cos(\omega t-\phi)| \tag{6}$$

A contains A' and a constant K to fit the function $|V^m|$ to $\exp(-b/|V^{1/2}|)$.

It should be stressed that the general results obtained are not sensitive to the particular value of m chosen for the exponent in this approximation. Any power between, say 2 and 8, will lead to substantially similar solutions from which the conclusions of the next section follow. The basic equation for our discussion remains (5), not the specific approximation used in equation (6) which only serves to facilitate calculations. Which power of V we select is only important if computed results are to be compared quantitatively to the brightness waves observed for specific panels.

Equation (6) can be integrated to give the steady-state solution

$$0 \leqq \omega t \leqq \frac{\pi}{2}+\phi \qquad B = AI_0V_0^m[F_m(\omega t)+C_m(\omega\tau,\phi)\cdot e^{-t/\tau}]\frac{1}{\sqrt{[1+(m+1)^2\omega^2\tau^2]}}$$

$$\frac{\pi}{2}+\phi \leqq \omega t \leqq \pi \qquad B = AI_0V_0^m[-F_m(\omega t)+C'_m(\omega t,\phi)\cdot e^{-t/\tau}]\frac{1}{\sqrt{[1+(m+1)^2\omega^2\tau^2]}} \tag{7}$$

Here we have used the following symbols:

$$F_m(\omega t) = \sin^m\omega t\cdot\cos(\omega t-[\phi+\phi^*])+\cos(\phi+\phi^*)+$$

$$\sum_{\nu=0}^{\infty}\frac{m(m-1)(m-2)\ldots(m-2\nu)(\omega\tau)^{2\nu+1}}{[1+(m-1)^2\omega^2\tau^2][1+(m-3)^2\omega^2\tau^2]\ldots[1+(m-2\nu-1)^2\omega^2\tau^2]^{1/2}}\sin^{m-2\nu-2}\omega t\cdot\cos(\omega t+a_\nu)$$

$$\tan\phi^* = (m+1)\omega\tau$$

$$\cot a_\nu = (m-2\nu-1)\omega\tau$$

$$C_m(\omega\tau, \phi) = \left\{2F_m\left(\frac{\pi}{2}+\phi\right)\cdot\exp\left(\frac{\pi}{2}+\phi\right)\Big/\omega\tau\right) - 2F_m(0)\cdot\exp(\pi/\omega\tau)\Big\}(\exp(\pi/\omega\tau)-1)^{-1}$$

$$C'_m(\omega\tau, \phi) = C_m + 2F_m\left(\frac{\pi}{2}+\phi\right)\cdot\exp\left(\frac{\pi}{2}+\phi\right)\Big/\omega\tau\right)$$

While F_m (ωt) represents essentially emission due to excitation during the cycle (its first term follows the excitation term more or less), the term $C_m \exp(-t/\tau)$ represents decay from excitation present at the beginning of the cycle.

Our panels closely approximated a sixth-power law of brightness vs. voltage from the limit of phototube sensitivity up to dielectric breakdown, i.e. from 100 to 400 V on 4-mil phosphor layers. Therefore, we chose the substitution of V^5 for the exponential when we computed numerical values of the integral of equation (5). With $I_0 \sim V_0$ we then have $B \sim V_0^6$, as experimentally observed on our panels.

Equation (7) is rather complicated, and it is tedious to compute the time dependence of brightness from it. Equation (6) was therefore solved with $m = 5$ on an analog computer and solutions were plotted as a function of time for values of the parameters $\omega\tau$ and ϕ. Fig. 1 shows a few representative solutions.

IV. RESULTS

Some immediate qualitative conclusions can be drawn from an examination of the curves.

(1) Large values of ϕ lead to brightness waves without secondary peaks.

(2) Large values of $\omega\tau$ tend to smooth out secondary peaks.

(3) Large values of $\omega\tau$ also lead to brightness waves superimposed on a d.c. light level.

These points are in agreement with experimental observations. Four phosphors, all prepared from zinc sulfide as described in Table 1, were used for our experiments. At 1000 c/s, they exhibited the brightness waves shown in Fig. 2. Values of ϕ which give a good match between experimental and calculated curves are indicated.

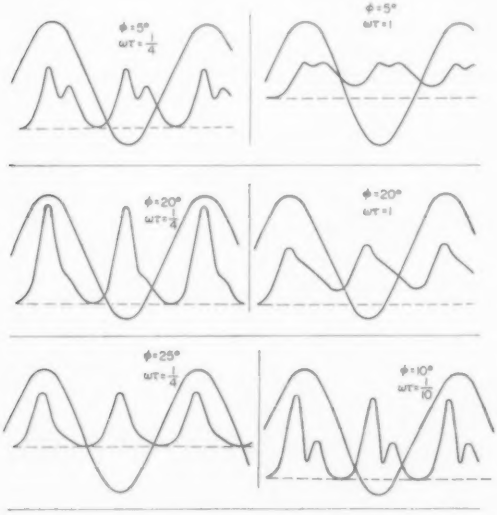


FIG. 1. Voltage wave form and calculated brightness waves for selected values of $\omega\tau$ and ϕ;– – – – zero light level.

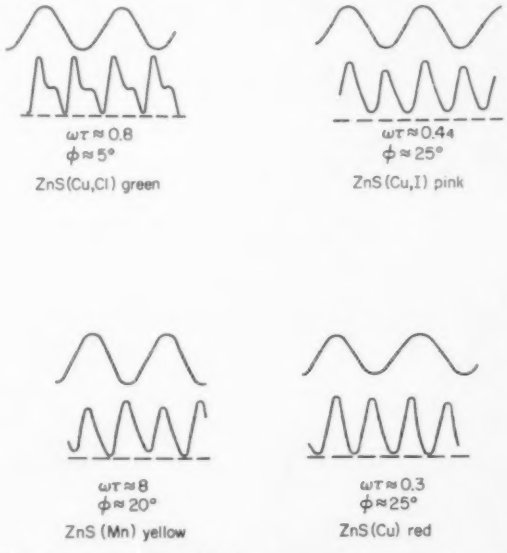


FIG. 2. Experimental brightness waves as a function of $\omega\tau$ and ϕ; – – – – zero light level.

Table 1. *Phosphor preparation*

Phosphor	Activation	Firing			Post firing
		Temp. (°C)	Time (hr)	Atm.	Treatment
Zn(S, O):(Cu, Cl)	0·005 g atom Cu/mol ZnS 0·02 g atom Cl/mol ZnS 0·4 mol ZnO/mol ZnS	900	24	air	A+B
ZnS:(Cu, I)	0·0006 g atom Cu/mol ZnS 0·067 g mol NH_4I/mol ZnS	1150	2	dry H_2S	B
ZnS: (Cu)	0·003 g atom Cu/mol ZnS	1100	2	dry H_2S	B
ZnS: (Mn, Cu)	0·0025 g atom Cu/mol ZnS 0·0125 g atom Mn/mol ZnS	1100	1	dry H_2S	B

A. Washed in hot acetic acid (20 per cent), rinsed with water.
B. Washed in NaCN solution (3 per cent), water, methanol. Dried for 16 hr in air at 125°C.

Fig. 3 shows brightness waves for a ZnS:(Cu, Cl) phosphor at 1 kc/s and 20 kc/s with the same applied voltage. The secondary peak has been much suppressed at 20 kc/s.

The connection between ϕ and the properties of the phosphors is quite complex. Some experiments were directed toward its understanding. However, no simple scheme based on an equivalent circuit model proved satisfactory. It is well known that although the electroluminescent panels appear to be good condensers, they are anything but simple capacitors. Fig. 4 shows a discharge curve of a ZnS:(Cu, Cl) panel. For comparison the discharge curve of an air-capacitor/carbon-deposited resistor taken under the same conditions is shown. Attempts to explain the behavior of our panels in terms of a series parallel *RC* network led to somewhat ambiguous relations between ϕ and the properties of the phosphors. We did find, however, that in a general way phosphors which have regular brightness waves are more lossy and have higher ϕ values than those which exhibit pronounced secondary peaks (see Fig. 2).

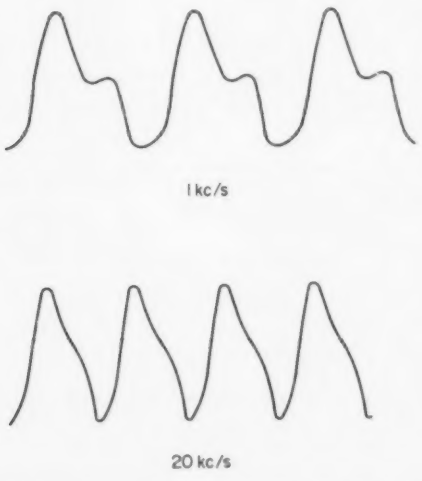


FIG. 3. Experimental brightness wave for ZnS:(Cu, Cl) panel at two frequencies.

The general solution of the differential equation (6) contains a transient term on the order of $\exp(-t/\tau)$ which must disappear before the steady state is reached. Fig. 5 shows a calculated plot of instantaneous brightness over the first few cycles indicating clearly the transient and its approach to steady state. An examination of equation (7) shows that the phase relationship of

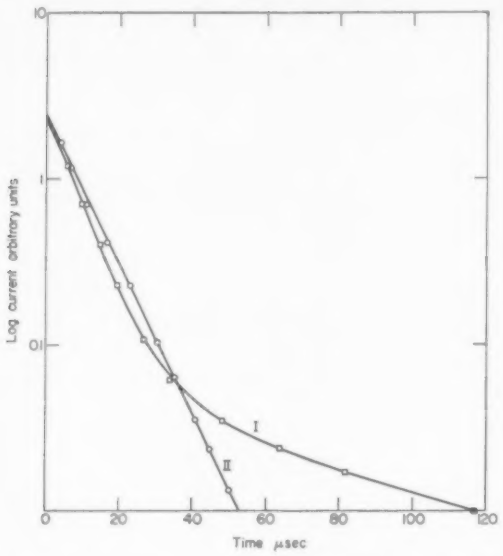


FIG. 4. Discharge curves of EL panel (I) and ideal R–C network (II).

the main brightness peak to the voltage peak is a function of $\omega\tau$. For constant τ, it changes from leading the voltage wave form to lagging it as frequency is increased. Conversely, as τ is increased, at constant frequency the same effect is created. This offers an explanation for the changes in phase which have been seen experimentally as a function of frequency and temperature.[1]

The average brightness of the panel can be defined as

$$\bar{B} = \frac{\int_0^{\pi/\omega} N_e/\tau \, dt}{\int_0^{\pi/\omega} dt} \tag{8}$$

If equation (7) is substituted into equation (8) and the integration carried out, one finds that

$$\bar{B} = AI_0V_0^m f(\phi) \tag{9}$$

where $f(\phi)$ is a slowly varying function of ϕ and for $m = 5$ has values shown in Table 2.

Table 2

ϕ (deg)	$f(\phi)$
0	1·000
5	1·019
10	1·075
20	1·284
30	1·587
45	2·102
90	2·945

The I_0 that appears in this equation is related to V_0 through an impedance term. If one assumes the impedance of a series resistor and capacitor, a very rough approximation only, then

$$I_0 \simeq \frac{V_0}{\sqrt{[R^2+(1/C^2\omega^2)]}} = \omega V_0 \frac{C}{\sqrt{(1+R^2C^2\omega^2)}} \tag{10}$$

For small values of ω, we thus obtain

$$\bar{B} \simeq \omega V_0^{m+1} f(\phi) \tag{11}$$

or, if we resubstitute the exponential term,

$$\bar{B} \simeq \omega V_0 \exp(-b/V_0^{1/2}) f(\phi) \tag{12}$$

This relation looks similar to empirical expressions for average brightness which have been proposed[14] except that frequency is explicitly contained in it. Of course, it is questionable since it

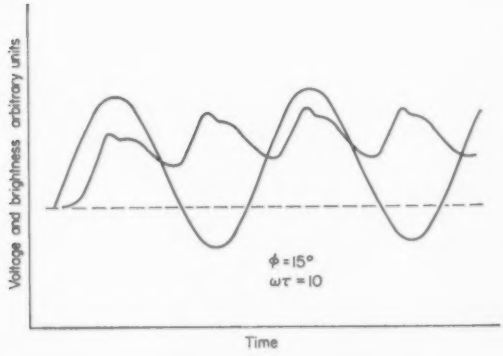


FIG. 5. Calculated rise in brightness during first few cycles of excitation; – – – – zero light level.

has only been derived for low frequencies, and since it depends on a simplified formulation of an equivalent circuit for the electroluminescent panel. Nevertheless, it is interesting and provocative that a series *RC* equivalent circuit predicts frequency behavior for average brightness which can be observed experimentally at low frequencies (essentially a linear dependence).

V. LIGHT DECAY CONSTANTS

The assumption of the monomolecular decay in brightness is only to a very rough approximation true in our phosphors, and indeed, in any phosphors. Many luminescent-center recombinations can be described as an initial monomolecular process changing into a polymolecular process. However, the appearance of secondary peaks does not depend on a specific assumption of monomolecular brightness decay used in equation (6). No assumptions about delayed recombinations due to trapping or other causes are necessary to account for secondary peaks in the brightness waves. Of course, this does not mean that there is no trapping in phosphors, it is simply pointed out that secondary peaks can be explained without reference to traps. In fact, delayed recombinations would seem to have a tendency to smooth out the secondary peaks, other things being equal. In principle, equation (5) can be modified to apply to processes of any complexity (bimolecular processes, trapping, etc.), although the mathematics rapidly get out of hand.

The steady-state average brightness is of course not dependent on the trapping or recombination scheme. At steady state, dN_e/dt is by definition 0 over a cycle, and the principle of detailed balance shows that at steady state any trapping process and its reverse proceed at the same over-all rate. Therefore, any mode of recombination will give us

$$\bar{B} = \int_0^1 \frac{N_e}{\tau} dt = \frac{A'}{\pi} \int_0^{\pi} |I| \times \exp(-b/|V^{1/2}|) d(\omega t) \text{ (per unit time)} \tag{13}$$

VI. CONCLUSIONS

The secondary peaks in sinusoidal brightness waves can be accounted for by considering an excitation term which depends on the voltage and a current component of an electroluminescent panel. No assumptions about delayed recombinations are necessary. A complete description of the critical parameter ϕ awaits a better understanding of dielectric properties of electroluminescent phosphors. A simple equivalent circuit model gives an equation for average brightness which contains frequency effects, and which can be observed qualitatively by experiment.

The approach taken in this paper is largely phenomenological in nature and it remains to be seen whether the theory presented here can be fitted into current detailed models about the mechanism of electroluminescence. The major difficulty appears in the derivation of the excitation term which is not strictly justifiable on firm theoretical grounds. A justification awaits the complete solution of the problem of transient behavior of space charges in semiconductors.

Acknowledgement—The authors would like to thank Mr. R. ZALLEN, now at the Gordon McKay Laboratory, Harvard University, Cambridge, Mass., for his assistance with some of the experiments and discussions of the manuscript.

REFERENCES

1. J. MATTLER, *J. Phys. Radium* **17,** 725 (1956).
2. F. F. MOREHEAD, *J. Electrochem. Soc.* **105,** 461 (1958).
3. P. ZALM, *Phillips Res. Rep.* **11,** 353 (1956).
4. G. F. ALFREY and J. B. TAYLOR, *Proc. Phys. Soc. Lond.* **B68,** 775 (1955).
5. F. MATOSSI, *Phys. Rev.* **98,** 434 (1955).
6. D. HAHN and F. W. SEEMANN, *Z. Naturf.* **10a,** 586 (1955).
7. W. THORNTON, *Phys. Rev.* **102,** 38 (1956).
8. C. H. HAAKE, *J. Appl. Phys.* **28,** 117 (1957).
9. W. W. PIPER and F. E. WILLIAMS, *Brit. J. Appl. Phys.* Suppl. 4, 39 (1955).
10. J. B. TAYLOR, *Brit. J. Appl. Phys.* Suppl. 4, 45 (1955).
11. F. SEITZ, *Phys. Rev.* **76,** 1376 (1949).
12. G. DESTRIAU, *Brit. J. Appl. Phys.* Suppl. 4, 49 (1955).
13. F. A. KRÖGER, *Naturwissenschaften* **29,** 80 (1956).
14. G. DESTRIAU, *Proc. I.R.E.* **43,** 1911 (1955).

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 18–23. Printed in Great Britain

FIELD-RETARDED DIFFUSION OF ANTIMONY IN GERMANIUM*

C. FA and R. ZULEEG

Hughes Aircraft Company, Semiconductor Division, Newport Beach, California

(*Received* 14 *November* 1960; *in revised form* 13 *February* 1961)

Abstract—Diffusion techniques, applied to semiconductors to produce impurity gradients, have revealed that under certain conditions Fick's law is not always obeyed. It will be shown experimentally and theoretically that a field-retarded diffusion of antimony will occur in germanium in the presence of a p-type impurity, e.g. gallium, which is higher in concentration and has a negative gradient in the direction of the donor diffusion. This implies, in addition, temperature settings whereby the intrinsic carrier concentration n_i is smaller than N_A, the number of ionized p-type impurity atoms. The investigations have been carried out for a temperature range of 700–900°C and with concentration levels of $N_A \leqslant 10^{20}$ atoms/cm^3 and of $N_D \leqslant 10^{19}$ atoms/cm^3.

By including field terms it is shown that the diffusion constant of antimony D_{Sb} is dependent upon the ratio of its concentration N_D to the intrinsic carrier concentration n_i, and the ratio of acceptor and donor gradients. The experimental evidence is in general agreement with the theoretical formulation.

Résumé—Certaines techniques de diffusion appliquées aux semiconducteurs pour produire des pentes d'impuretés ont révélé que, sous certaines conditions, la loi de Fick n'est pas toujours obéie. Il sera démontré expérimentalement et théoriquement que la diffusion, retardée par le champ de l'antimoine, se produira dans le germanium en la présence d'une impureté du type p, comme par exemple le gallium qui a une plus forte concentration et qui possède une pente négative dans la direction de la diffusion des donneurs. Ceci demande aussi des ajustements de température ou la concentration intrinsèque des porteurs n_i est plus petite que N_A, le nombre d'atomes d'impureté ionisés de type p. Les expériences ont été faites dans la gamme de température de 700°C à 900°C et avec des niveaux de concentration de $N_A \leqslant 10^{20}$ atomes/cm^3 et $N_D \leqslant 10^{9}$ atomes/cm^3.

En comprenant les termes du champ, il est demontré que la constante de diffusion de l'antimoine, D_{Sb}, dépend du rapport de sa concentration N_D à la concentration intrinsèque des porteurs n_i, et le rapport des pentes d'accepteurs et de donneurs. Les résultats expérimentaux sont en accord avec les formules théoriques.

Zusammenfassung—Diffusionsprozesse werden häufig angewendet um Störstellengradienten in Halbleitern zu erzeugen. Es hat sich jedoch gezeigt, dass Fick's Gesetz der Diffusion unter bestimmten Bedingungen nicht immer befolgt wird. In dieser Arbeit wird experimentell nachgewiesen, dass eine feldverzögerte Diffusion von Antimon in Germanium auftreten kann in der Gegenwart einer p-leitenden Verunreinigung, z.B. Gallium. Um die verzögerte Diffusion zu bewirken wird verlangt, dass die Galliumkonzentration grösser ist als diejenige von Antimon und dass ein negativer Gradient in der Richtung der Antimondiffusion vorhanden ist. Der Temperaturbereich in dem diese Verzögerung auftreten kann ist beschränkt und bedingt eine Inversionsdichte n_i kleiner als N_A, die Anzahl der ionisierten, p-leitenden Verunreinigungsatome.

Die Untersuchungen werden in einem Temperaturbereich von 700°C bis 900°C und in einem Verunreinigungsbereich von $N_A \leqslant 10^{20}$ Atome/cm^3 und $N_D \leqslant 10^{19}$ Atome/cm^3 durchgeführt.

Es wird theoretisch gezeigt, dass bei einer Addierung eines Feldstromes zum Diffusionstrom die Diffusionskonstante von Antimon D_{Sb} eine Form annimmt, die abhängig ist von (a) dem Verhaeltnis der Konzentration N_D zu der Inversionsdichte n_i und (b) dem Verhaeltnis der Akzeptoren- und Donatorengradienten. Die Ergebnisse des Experimentes sind in allgemeiner Uebereinstimmung mit der theoretischen Begründung.

* Presented at the IRE–AIEE Solid State Device Research Conference, 13–15 June 1960, Pittsburgh, Pennsylvania.

INTRODUCTION

IT HAS been a very common assumption that FICK's law of diffusion[1] can be applied to solid-state diffusion in semiconductors. This law states that the diffusion rate J depends entirely on the impurity gradient, i.e.

$$J = -D\,dc/dx \qquad (1)$$

where D is the diffusion constant of the diffusant and c its concentration. The diffusion constant in FICK's law is inherent for a given material and is only temperature dependent. It has been customary to describe diffusion in semiconductors with a diffusion constant which has only the temperature dependence. In recent theoretical and experimental investigations,[2–6] it has been shown that, under certain diffusion conditions, D is not a constant of the material dependent on the temperature alone; electrostatic and electrodynamic forces may contribute to considerable deviation.

Electrostatic interaction, such as described in Refs. 3 and 5, will certainly decrease with increasing temperature and impurity concentration and, hence, this interaction can be eliminated from the results reported, especially since the intrinsic number of holes or electrons will be less than the number of ionized impurity atoms. In this latter case, it was shown[2,3] that field-aided diffusion can occur, which increases the diffusion constant by a factor of 2. The diffusion constant is, therefore, not a constant at certain temperatures for a particular diffusant, but depends also on impurity concentration. The actual field arises from the action of holes associated with the ionized p-type impurity atoms. The holes would ordinarily tend to diffuse away from the impurity ions by virtue of large differences in diffusion constants. This tends to set up a space charge; in order to prevent this and to keep the two distributions identical, a field will be established. The influence of this field on diffusion properties has been studied in a system Ga–Sb–Ge in the temperature range from 700 to 900°C for various impurity concentrations and gradients. It will be shown, in the special case of germanium, where most of the n-type impurities of the fifth column diffuse faster by several orders of magnitude than the p-type impurities, that this field introduces forces upon the diffusion of the faster-diffusion n-type impurities which tend to retard the diffusion if the gradients of donors and acceptors are in the same direction. In the opposite case, a field-aided diffusion would result. Finally, a boundary condition is realized whereby diffusion ceases completely owing to the field-opposing forces.

FICK's law, therefore, is no longer valid and must include terms describing field-aided or retarded diffusion.

EXPERIMENTAL PROCEDURE

A germanium crystal of p-type conductivity with an acceptor doping level around 10^{16} atoms/cm^3 has been used as a starting material. Slices of thickness 25×10^{-3} cm have been lapped and polished to obtain smooth surfaces for even indiffusion. Since shallow diffusion depths have been chosen for the experiments, the work damage remaining after polishing was removed by a chemical polish.

Part of the slices was subjected to p-type indiffusion of gallium in an open-tube, flowing-gas furnace at a temperature of 850°C for 2 hr. A p^+-indiffusion gradient with a characteristic diffusion length, $L = 2\sqrt{(Dt)}$, of $1{\cdot}0\,\mu$ and a surface concentration, $C_{0\,\mathrm{Ga}}$, equal to 10^{20} atoms/cm^3 resulted from this treatment. An open-tube, flowing-gas, two-zone furnace has been used to perform indiffusion of antimony from the vapor phase at a constant source temperature. Slices with and without p^+-diffusion from the same material were then simultaneously indiffused with antimony. In the evaluation, the temperature of the diffusion and the surface concentration of antimony served as variables for the experiment.

After the n-type diffusion cycle, angle lapping is used in conjunction with copper-staining techniques to reveal the p–n junction in the p-type material and the p–n–p junction structure in the previously p^+-diffused material. (See Figs. 1–4.)

Sheet resistivity measurements on the p-type material with the formed p–n-junction by the antimony diffusion, together with the values of the junction depth x_{p0}, have been used by employing BACKENSTOSS' method[7] to determine the diffusion constant of Sb as a function of temperature. These values are confirmed by diffusion-constant measurements of Sb reported by BLAKESLEE[8]. Since this evaluation was based upon a solution of FICK's law according to equation (1), the antimony diffusion up to 10^{19} atoms/cm^3 seems to follow

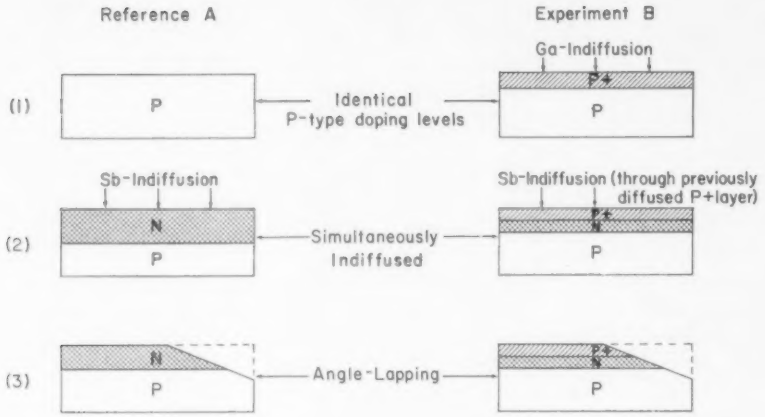


FIG. 1. Principal steps 1–5 of the experiment to produce specimens.

FICK's law within the error limits of the experiment. The error function solution for the boundary condition of a constant-concentration diffusion source is, therefore, appropriate in the form

$$C(x) = C_0 \operatorname{erfc}[x/2\sqrt{(Dt)}] \qquad (2)$$

where C_0 is the surface concentration, $C(x)$ the concentration at a distance x from the surface, t the duration of the diffusion cycle and D the diffusion constant which is determined by the diffusant and the medium through which it diffuses.

In our evaluation procedure and mainly for reasons of a comparison, the assumption which is somewhat objectionable according to Ref. 5 has been made that the donor surface concentration, $C_{0\,Sb}$, during the simultaneous diffusion into the p-type and previously p^+-diffused material is the same. Under identical diffusion conditions, consistent deviations in the junction-formation distances, x_{p2} from x_{p0}, have been observed. x_{p2} and x_{p0} are the depths of the junctions formed by the Sb diffusion in the p-type material and the previously p^+-diffused material respectively. Therefore, it was obvious that the n-type diffusion in presence of the p-type gradient does not follow FICK's law of diffusion, and additional forces must be responsible for the difference. For a particular set of conditions, namely $C_{0\,Ga} = 10^{20}$ atoms/cm^3, and $C_{0\,Sb} = 10^{19}$ atoms/cm^3 and $t = 30$ min, the temperature range of 700–900°C has been chosen to study the diffusion behavior of antimony in presence of the p-type gradient. For the purpose of evaluation, an apparent diffusion constant D_{Sb} has been defined by

$$D_{Sb}(\text{apparent}) = 1/4t[x_{p2}/\arg \operatorname{erfc}(C_0/C_{0\,Sb})]^2 \qquad (3)$$

If a junction in the p-material has been formed at x_{p0} under the conditions of

$$C_0 = C_{0\,Sb} \operatorname{erfc}[x_{p0}/2\sqrt{(D_{Sb}t)}] \qquad (4)$$

substitution of (4) into (3) then yields for equal times t

$$D_{Sb}(\text{apparent}) = D_{Sb}[x_{p2}/x_{p0}]^2 \qquad (5)$$

Figs. 5 and 6 present the experimental results obtained by using equation (5) and measuring the junction depths, x_{p0} and x_{p2}, formed at various temperatures in the p-type and previously p-type indiffused samples.

THEORETICAL DIFFUSION CONSIDERATIONS

An excellent and extensive theoretical treatment

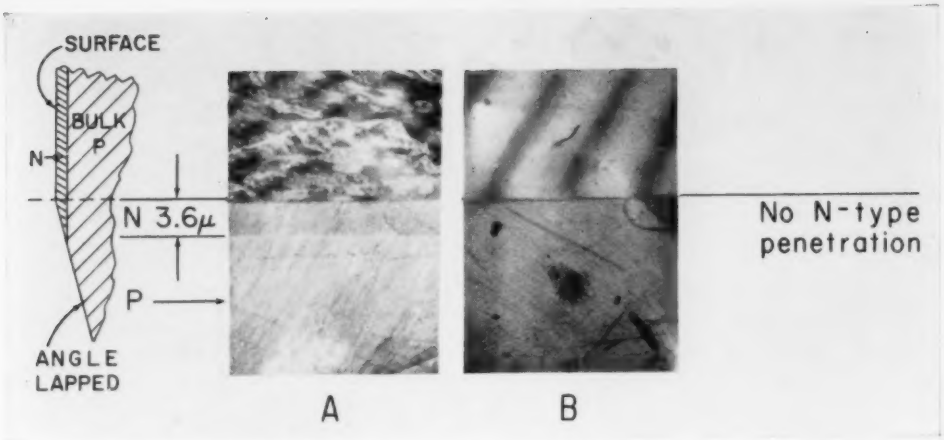


FIG. 2. Angle-sections of specimens at 800°C. A. Single Sb diffusion at 800°C for 30 min. $N_{0D} = 2 \times 10^{19}$ atoms/cm^3; B. same Sb diffusion through Ga diffused layer of $N_{0A} = 10^{20}$ atoms/cm^3. Bulk material: Ge with $N_A \simeq 10^{16}$ atoms/cm^3.

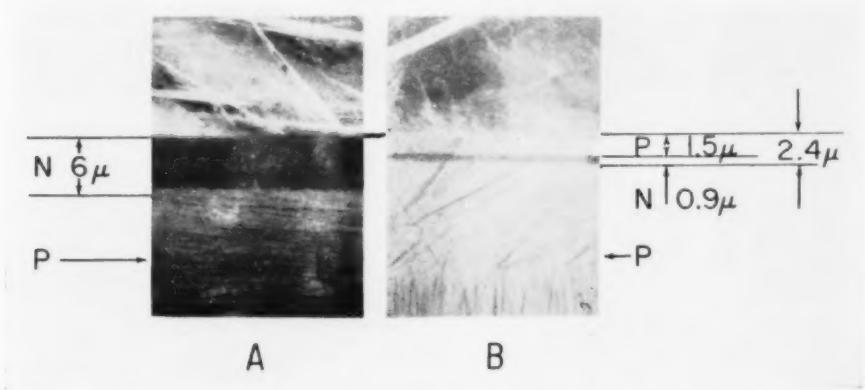


FIG. 3. Angle-sections of specimens at 825°C. A. Single Sb diffusion at 825°C for 30 min. $N_{0D} = 3 \times 10^{18}$ atoms/cm^3; B. same Sb diffusion through Ga diffused layer of $N_{0A} = 10^{20}$ atoms/cm^3. Bulk material: Ge with $N_A \simeq 10^{16}$ atoms/cm^3.

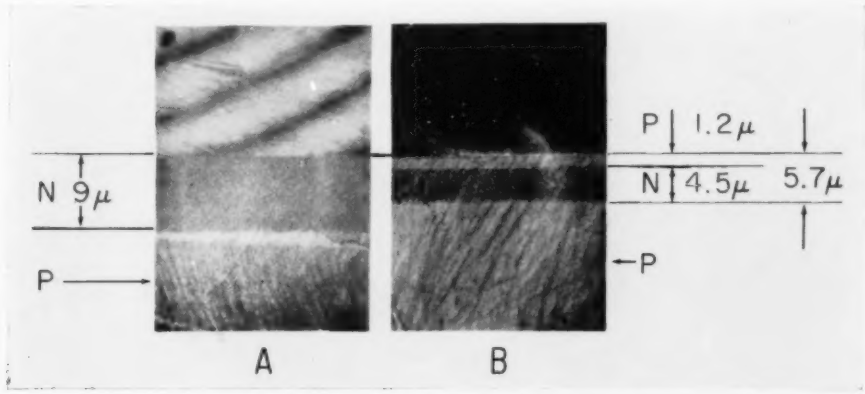


FIG. 4. Angle-sections of specimens at 850°C. A. Single Sb diffusion at 850°C for 30 min. $N_{0D} = 3 \times 10^{18}$ atoms/cm^3; B. same Sb diffusion through Ga diffused layer of $N_{0A} = 10^{20}$ atoms/cm^3. Bulk material: Ge with $N_A \simeq 10^{16}$ atoms/cm^3.

facing p. 20

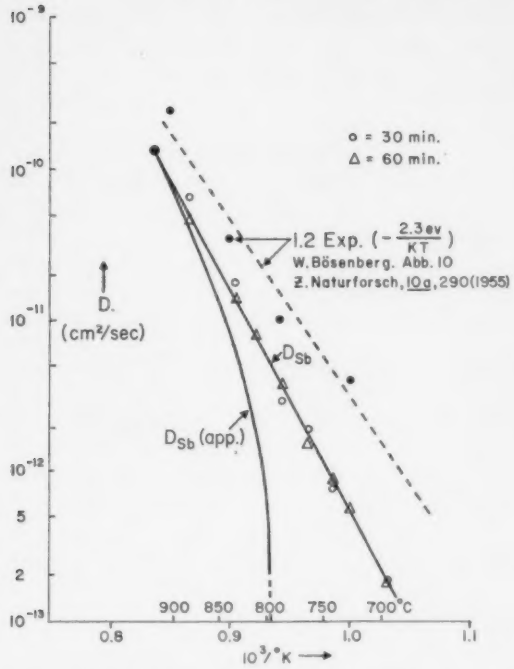


FIG. 5. Plot of diffusion constant of antimony vs. temperature from our determination in the retarded and normal diffusion case. The given values of D_{Sb} (apparent) apply for concentration levels of $N_{A0} = 10^{20}$ atoms/cm^3 and $N_{D0} = 10^{19}$ atoms/cm^3 and a characteristic diffusion length of $2\sqrt{(Dt)} = 1\ \mu$ for gallium.

of diffusion problems in semiconductors by including field terms has been given by ZAROMB[9].

According to equation (18) of this publication, we can write our donor diffusion current in the form

$$J_D = -D_D\, dN_D/dx - \mu_D N_D E \qquad (6)$$

where μ_D is the field mobility of the diffusing particle and $-E$ is the average electrical force acting on a positively charged particle. The field E which is acting upon the diffusing antimony particles in our experiment can be found from equation (32) of Ref. 9. Since the diffusion constants of holes and electrons, D_h and D_e respectively, are much larger than the diffusion constants of acceptors and donors, D_A and D_D respectively, and in addition $N_A \gg N_D$ in our experiment, it is easily shown that

$$E \simeq -\frac{D_D}{2n_i\mu_D}\, dN_A/dx \qquad (7)$$

The expression for the field E in equation (7) can now be inserted into equation (6) and will result in a form describing the diffusion of the faster-moving impurity species. Normally, the slow-moving acceptor particles would have to be considered too, when the net impurity distribution at various distances is under investigation. Aided diffusion of acceptor impurity atoms in our case has been neglected and ruled out by the large difference in the diffusion constants of antimony and gallium. FICK's modified law for the antimony

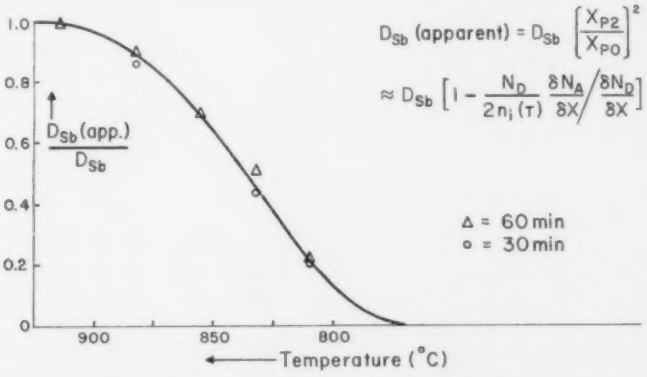


FIG. 6. Dependence of the retarded diffusion of antimony on time and temperature for $N_{A0} = 10^{20}$ atoms/cm^3 and $N_{D0} = 10^{19}$ atoms/cm^3.

diffusion will now read

$$J_{D(\mathrm{Sb})} = -D_{D(\mathrm{Sb})} dN_{D(\mathrm{Sb})}/dx \times \left[1 - \frac{N_{D(\mathrm{Sb})}}{2n_i}\left(\frac{dN_{A(\mathrm{Ga})}}{dx} \Big/ \frac{dN_{D(\mathrm{Sb})}}{dx}\right)\right] \quad (8)$$

The subscripts Sb and Ga are used in reference to antimony and gallium. Equation (8) is the approximate solution which will be used to interpret the reported diffusion of antimony in germanium in the presence of gallium. The solution (8) is in general applicable only to a system whereby $N_A > N_D$ and $D_D \gg D_A$.

APPLICATION TO EXPERIMENT

In Fig. 5, the measured diffusion constant of antimony is plotted as a function of temperature. Previous measurements are indicated and are somewhat higher than ours. Recently published data[8], however, do agree with the values reported. Retarded antimony diffusion constant, D_{Sb} (apparent), for our specific experimental arrangement is shown on the same plot. It should be pointed out that this retarded diffusion is subject to changes in the original gallium impurity concentration and its impurity gradient. At the lower end of the temperature range of our investigation, the diffusion constant of antimony departs by several orders of magnitude from the normal one and its behavior can be described by equation (8). At the higher end of the temperature range and finally approaching the melting point of germanium, quite normal diffusion of antimony will occur according to FICK's law [equation (1)]. Transition of equation (8), which describes the retarded diffusion into equation (1) in the form

$$J_{D(\mathrm{Sb})} = -D_{D(\mathrm{Sb})}\, dN_{D(\mathrm{Sb})}/dx \quad (9)$$

which describes normal diffusion according to FICK's law, is caused in the first-order effect by the temperature dependence of the intrinsic carrier concentration n_i. The spacial impurity gradient $dN_{A\,(\mathrm{Ga})}/dx$, of course, is a characteristic variable in our experiment and has been kept constant in regard to $dN_{D\,(\mathrm{Sb})}/dx$. For other diffusion parameters, both impurity gradients can play different roles; and, in cases where the gradient of one impurity species is zero, while that of the other is high, these terms predict counter-diffusion of the former species in the direction of increasing concentration. The interaction between arsenic and aluminum in germanium[4] is an example of such counter-diffusion.

To show consistency of the retarding effect, diffusion runs were carried out at different time cycles, but with identical temperature settings. The results are demonstrated in Fig. 6.

To check the quantitative agreement of equation (8) with the experimental values of Fig. 5, the example is restricted to the case where

$$(dN_A/dx)/(dN_D/dx) = 1$$

Assuming that $N_{D\,(\mathrm{Sb})} = 2 \times 10^{19}$ atoms/cm^3, one finds for n_i at 800°C a value of approximately 10^{19} atoms/cm^3,[5] and $D_{D\,(\mathrm{Sb})}$ tends to take on a very small value. On the other hand, for a temperature of 850°C, where $N_{D\,(\mathrm{Sb})} \simeq n_i$, one finds $D_{D\,(\mathrm{Sb})}$ (apparent) $= \frac{1}{2} D_{D\,(\mathrm{Sb})}$. These two theoretical points are in fair agreement with the experimental points in Fig. 5.

SUMMARY

It has been shown experimentally that, under certain conditions, a field-retarded diffusion can exist in semiconductors. The conditions in our case are such that the gallium (acceptor) surface concentration is larger than the antimony (donor) surface concentration during the diffusion. Antimony as a donor indiffusion will then be retarded in presence of a negative impurity gradient of gallium if its concentration is larger than the intrinsic carrier concentration. The retarding effect will disappear as soon as the antimony concentration is compatible with or even smaller than the intrinsic carrier concentration. The impurity gradient of the stationary acceptors and the varying gradients of the diffusing donors will be responsible for the additional driving forces acting upon the mobile impurity atoms. By including field terms, a modified FICK's law describes the retarded diffusion of antimony observed.

Acknowledgements—The authors wish to acknowledge the encouragement and technical advice of Mr. W. P. WATERS in preparing this paper. Mr. M. SIRACUSA is credited for performing most of the diffusion experiments and evaluation procedures.

REFERENCES

1. A. Fick, *Ann. Phys., Lpz.* **94,** 59 (1855).
2. A. D. Kurtz and R. Yee, *J. Appl. Phys.* **31,** 303 (1960).
3. A. D. Kurtz and C. L. Gravel, *J. Appl. Phys.* **29,** 1456 (1958).
4. J. O. McCaldin, *J. Appl. Phys.* **31,** 89 (1960).
5. H. Reiss, C. S. Fuller and I. D. Morin, *Bell Syst. Tech. J.* **35,** 535 (1956).
6. F. M. Smits, *Proc. I.R.E.* **46,** 1049 (1958).
7. G. Backenstoss, *Bell Syst. Tech. J.* **37,** 699 (1958).
8. A. E. Blakeslee, *Open Tube Diffusion of Antimony into Germanium.* Paper presented at Electrochemical Society Meeting, Chicago, May 1–5, 1960.
9. S. Zaromb, *IBM J. Res. Dev.* **1,** 57 (1957).

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 24–28. Printed in Great Britain

PHOTOCONDUCTIVITY IN SINGLE-CRYSTAL TELLURIUM†

D. F. EDWARDS,‡§ C. D. BUTTER¶ and L. D. McGLAUCHLIN¶

(*Received* 9 *January* 1961; *in revised form* 28 *February* 1961)

Abstract—Measurements are reported of the photoconductive properties of tellurium crystals grown from the vapor phase as well as from the melt. The results are compared with other types of photoconductive cells and with theoretical estimates of the cell detectivity. For the "best" tellurium cell the measured detectivity at the spectral peak was D_λ^* $(3{\cdot}4\,\mu, 900) = 2{\cdot}3 \times 10^{11}$ cm $(c/s)^{1/2}$ W^{-1} compared with the estimated value of D_λ^* $(3{\cdot}4\,\mu, —) = 2{\cdot}2 \times 10^{11}$ cm $(c/s)^{1/2}$ W^{-1}. This demonstrates that it is possible to construct tellurium cells whose detectivity is limited by background radiation.

Résumé—On rapporte les mesures des propriétés photoconductives des cristaux de tellurium développés de la phase vapeur aussi bien que de la fonte. Les résultats sont comparés avec d'autres types de cellules photoconductives et avec des évaluations théoriques de la "détectivité" de la cellule. Pour la meilleure cellule de tellurium, la "détectivité" mesurée à la pointe spectrale était de D_λ^* $(3{,}4\mu, 9000) = 2{,}3 \times 10^{11}$ cm $(c/s)^{1/2}W^{-1}$ comparée avec une valeur estimée de $D_\lambda^*(3{,}4\mu, —) = 2{,}2 \times 10^{11}$cm $(c/s)^{1/2}$ W^{-1}. Ceci démontre qu'il est possible de produire des cellules de tellurium dont la "détectivité" est limitée par la radiation de fond.

Zusammenfassung—Die Photoleitfähigkeit von Tellurkristallen wurde gemessen, wobei die Kristalle sowohl aus der Dampfphase als auch aus der Schmelze gewonnen wurden. Die Ergebnisse wurden mit anderen Typen von photoleitenden Zellen und mit den theoretischen Werten der Nachweisempfindlichkeit verglichen. Für die ,,besten" Tellurzellen ergab die Messung der Nachweisempfindlichkeit am spektralen Spitzenwert $D_\lambda^*(3{,}4\mu, 9000) = 2{,}3 \times 10^{11}$ cm $Hz^{1/2}/W$ gegenüber dem theoretischen Wert $D_\lambda^*(3{,}4\mu, —) = 2{,}2 \times 10^{11}$ cm $Hz^{1/2}/W$. Daraus ergibt sich die Möglichkeit der Herstellung von Tellurzellen, deren Nachweisempfindlichkeit durch Untergrundstrahlung begrenzt ist.

TELLURIUM single crystals grown from the vapor phase as well as from the melt display photoconductive effects comparable to lead sulfide photoconductive cells. In the present paper, measurements are reported for several tellurium photoconductive cells and the results are compared with other types of photoconductive cells, and with theoretical estimates of the cell characteristics.

The use of tellurium as a photoconductive element was first investigated by MOSS[1] who used evaporated thin films. For a cell cooled to liquid-nitrogen temperature, he reported a detectivity‖ value of D_λ^* $(1\,\mu, 80) = 1{\cdot}7 \times 10^9$ cm $(c/s)^{1/2}$ W^{-1}, and a response time of about 550 μsec for the cell having the greatest detectivity. This same tellurium film had only a moderate detectivity [D_λ^* $(1\,\mu, 80) = 10^5$–10^6 cm $c/s^{1/2}$ W^{-1}] at room temperature. LOFERSKI[2] later made photoconductivity measurements on cooled single crystals of tellurium with about the same results

† This work was conducted by Project MICHIGAN under Department of the Army Contract (DA-36-039-78801) administered by the U.S. Army Signal Corps.

‡ Institute of Science and Technology, The University of Michigan, Ann Arbor, Michigan.

§ Present address: Massachusetts Institute of Technology, Lincoln Laboratory, Lexington, Massachusetts.

¶ Honeywell Research Center, Hopkins, Minnesota.

‖ The conditions of measurement of the detectivity are specified by writing D_λ^* $(1\,\mu, 80\ c/s)$ where the asterisk indicates that the cell area and amplifier bandwidth have been normalized to 1 cm^2 and 1 c/s, respectively; the subscript λ indicates that this is the monochromatic detectivity for the wavelength specified by the first number in parenthesis. The second number in the parenthesis is the chopping frequency of the radiation. If a 500°K black body is used as the radiation source, the λ subscript is omitted and the 1 μ is replaced by 500°K, e.g., D^* (500°K, 80 c/s).

as Moss had found for the cooled evaporated films. The single crystals as used by LOFERSKI were cleaved from large tellurium crystals and etched to the desired dimensions. Nickel was electroplated to the ends of the samples before soldering the copper leads. The impurity concentration of these crystals was estimated by LOFERSKI to be about 10^{15} cm^{-3}.

SUITS[3] in 1957 found that thin hexagonal crystal prisms of tellurium when cooled to liquid-nitrogen temperature had detectivities and response times comparable to cooled PbS cells. These thin single crystal prisms (about $10 \times \frac{1}{2} \times \frac{1}{2}$ mm) were grown from the vapor phase in a low-pressure hydrogen atmosphere starting with 99·999 + per cent pure material,† and always resulted in a needle-like cell element. The electrical contacts were made by welding a hot wire, in most cases platinum, to the tellurium. The crystal was then mounted in a glass dewar shown in Fig. 1. Tellurium cells‡ having large detectivities have also been made from samples cleaved from melt-grown tellurium single crystals. Cells with sensitive elements as large as 16 mm^2 have been made in this way. The single crystals grown along the *c*-axis[5] by the Czochralski method had a high degree of perfection[6] as indicated by the low etch pit counts (as low as 100 cm^{-2}) and long carrier lifetime ($\sim$70 μsec at 300°K). BLAKEMORE *et al.*[6] report a change in room-temperature lifetime from 70 to about 1 μsec for an increase in the defect density from 10^4 to 10^6 cm^{-2}. Thus, the carrier lifetime is sensitively dependent on the defect density. The highest lifetime value previously reported[7] was 0·01 μsec, probably indicating a large defect density, (about 10^8–10^9 cm^{-2}).

Since tellurium can be plastically deformed under relatively small stresses, caution must be taken in preparing the tellurium samples to preserve the high degree of perfection. This is true for the vapor-phase-grown crystals as well as for those grown from the melt. As a convenience in referring to these two types of Te cells, the cells made with vapor-phase-grown elements will be given the suffix -*N* (e.g. Te-*N* or Te-59-*N*), meaning *needle-like*, and the suffix -*C* (e.g. Te-*C* or Te-100-*C*) for cells with elements *cleaved* from large single crystals. For the elements used in the Te-*C* cells, the electrical contacts were made of evaporated gold to which the leads were soldered. This method was adopted for these larger-area elements to reduce the effects of current-density variation across the element. The gold contacts also helped to define the sensitive area. Fig. 2 is an end view of a mounted tellurium element which was cleaved from a melt-grown crystal and has gold evaporated electrodes. The sensitive area is about 2×2 mm.

The electrical noise, 500°K black-body detectivity and spectral detectivity were measured for a number of tellurium cells. For these measurements the Te-*N* cells were connected directly across the primary of a G-5 Geoformer§ with impedance approximately matched to the cell dark resistance. For the Te-*C* cells, a low-noise transistor preamplifier was used. At optimum bias, the noise of the Te-*C* cells was considerably greater than the preamplifier noise. Fig. 3 shows typical noise-voltage values in a 5-cycle bandwidth plotted against

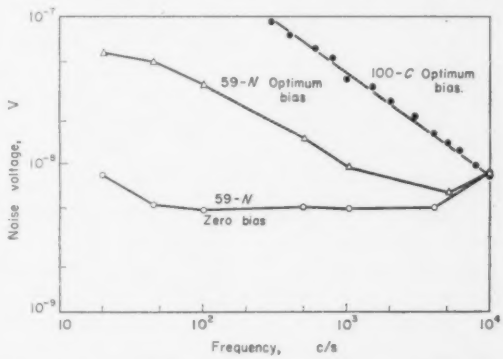


FIG. 3. Noise-voltage spectra for two types of tellurium cells.

frequency for both types of cells. For cell Te-59-*N* the noise voltage is given for both the optimum bias and zero bias. For this latter case the noise voltage is approximately frequency independent and equal to the thermal noise of the cell (2 mμV for 200 Ω at 77°K and 5-cycle bandwidth). For

† American Smelting and Refining Company, South Plainsfield, New Jersey.

‡ A preliminary report of Te cells utilizing large single crystals was presented by BUTTER and MCGLAUCHLIN at the Los Angeles Infrared Information Symposium.[4]

§ Triad Transformer Corporation, Venice, California.

optimum bias, i.e. the applied bias that produces maximum signal-to-noise ratio, the noise voltage for Te-59-N and Te-100-C are both typical semiconductor noise, i.e. the noise power has a $1/f$ frequency dependence. The increased noise at 10 kc/s for cell Te-59-N with and without bias is produced by the increased transformer losses at high frequencies.

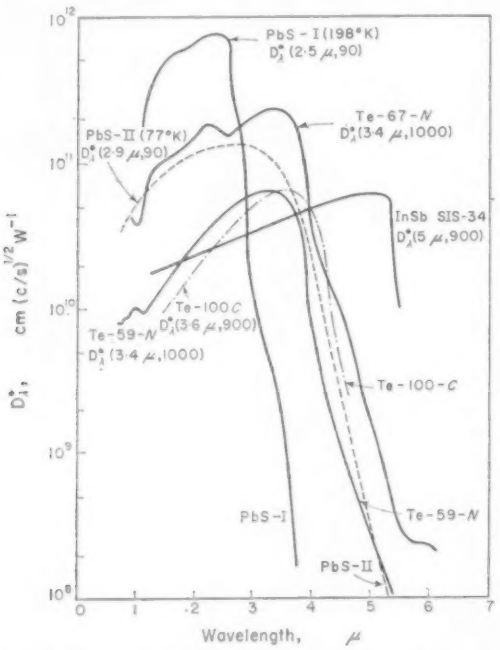


FIG. 4. Spectral detectivity for tellurium, indium antimonide and lead sulfide photoconductive cells.

The spectral detectivity D^*_λ is shown in Fig. 4 for typical Te-N and Te-C type cells. Also shown for comparison are curves for a few typical cells of different materials having D^*_λ curves competitive with those of the Te cells. Two curves are shown for the lead sulfide cells, PbS-I and PbS-II, to illustrate the two variations possible for this material. The PbS-I is the standard Eastman Type-N cell and PbS-II is the Eastman plumbide variation called Type-P. For PbS cells of the type represented by curve PbS-I, D^*_λ at the spectral peak is greater at 198°K than at 77°K, and has a long-wavelength cut-off of about 3·5 μ at 77°K. For cells of the type PbS-II, D^*_λ is greater at 77°K than at 198°K, and has a long-wavelength cut-off of about 4·5 μ at 77°K, and in general D^*_λ at the spectral peak is greater for the PbS-I type cells than for the PbS-II cells. The temperature effect for the two types of PbS cells has also been observed by SPENCER[8] who attributes the effect to the rate of absorption of ambient photons at the two temperatures.

The curve for the InSb cell shown in Fig. 4 is for a photoconductive cell reported by BRATT *et al.*[9] The cell temperature was 77°K and a cooled aperture was used to restrict the field of view to a 60° cone. The effect of the cooled aperture is to increase D^*_λ by about 1·5 times. ROBERTS[10] has reported D^*_λ values as great as 10^{11} cm(c/s)$^{1/2}$ W^{-1} at $\lambda = 5{\cdot}3\ \mu$ for InSb photoconductive cells of similar design.

From Fig. 4 it can be seen that cells Te-59-N and Te-100-C have very similar D^*_λ curves. The difference in the spectral peaks for these two cells is a surface effect as pointed out by LOFERSKI[2]. The long-wavelength edge for Te-100-C and Te-59-N is almost identical with that of cell PbS-II. At $\lambda = 3{\cdot}4\ \mu$ the D^* value for PbS-II is about 1·5 times that of the Te cells and increases for shorter wavelengths. The time constant for the Te cells is about 60–120 μsec compared to about 4000 μsec for the PbS-II cell. Thus, for applications where the shorter time constant is required and the increased D^*_λ for $\lambda < 3\ \mu$ is unimportant, the Te cells should be superior to the PbS-II cell. The D^*_λ values for the InSb cell Fig. 4, are greater than those for the Te cell except for a narrow spectral region centered about 3·6 μ. The time constant for these InSb cells is of the order of 1 μsec. Thus the InSb cells appear to be superior to the Te cells in applications where rapid response times are important. By placing an optical filter having the long-wavelength response of the Te cells in front of an InSb cell, one would have the advantages of the short time constant of the InSb cell and the spectral detectivity of the Te cell with little loss in D^*_λ at $\lambda = 3{\cdot}6\ \mu$.

Also shown in Fig. 4 is the D^*_λ curve for the tellurium cell having the greatest detectivity measured to date; it is labeled Te-67-N. At the spectral peak, 3·4 μ, the D^*_λ for Te-67-N is about 4 times that for Te-59-N and Te-100-C. The noise spectrum for cell Te-67-N is very similar to

FIG. 1. Glass dewar with sapphire window used for photoconductivity measurements on tellurium.

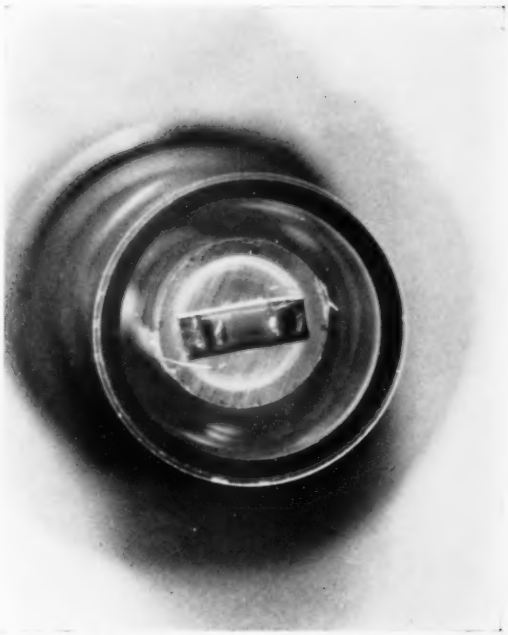


FIG. 2. Tellurium element cleaved from melt-grown crystal mounted in dewar.

[facing p. 26

that for cell Te-59-*N* (Fig. 3). Without an applied bias, the noise is approximately frequency independent and about equal to the thermal noise of the cell (1 mμV for 96 Ω at 77°K and 5-cycle bandwidth). With an applied bias the noise power is $1/f$ semiconductor noise. The performance characteristics are given in Table 1 for the three tellurium cells, Te-100-*C*, Te-59-*N* and Te-67-*N*.

Table 1. *Characteristics of tellurium single-crystal photoconductive cells*, 77°K

	Te-100-*C*	Te-59-*N*	Te-67-*N*
D^* (500°K, 90) [cm(c/s)$^{1/2}$W^{-1}]	$1{\cdot}3\times10^{9}$	$1{\cdot}3\times10^{9}$	$6{\cdot}2\times10^{9}$
D^* (500°K, 900) [cm(cs/$^{1/2}$W^{-1}]	4×10^{9}	$4{\cdot}7\times10^{9}$	$1{\cdot}7\times10^{10}$
D^*_λ (λ, 90) [cm(c/s)$^{1/2}$W^{-1}]	2×10^{10}	$1{\cdot}7\times10^{10}$	$8{\cdot}3\times10^{10}$
D^*_λ (λ, 900) [cm(c/s)$^{1/2}$W^{-1}]	$6{\cdot}4\times10^{10}$	$6{\cdot}2\times10^{10}$	$2{\cdot}3\times10^{11}$
λ (spectral peal) (μ)	3·6	3·4	3·4
τ (time constant) (μsec)	~60	~120	~120
R (resistance) (Ω)	1000	199	96
A (area) (cm^2)	$0{\cdot}02\times0{\cdot}02$	$0{\cdot}05\times0{\cdot}05$	$0{\cdot}05\times0{\cdot}05$

Recently, several articles have been published[11–13] that describe methods for calculating the ultimate detectivity of a cell on the assumption that the limitations are fluctuations of the background radiation. PETRITZ[11] has simplified the problem to a relation between the intrinsic energy gap E_t and the spectral detectivity $D^*_{\lambda_t}$. Taking the optical activation energy to be 0·37 eV as given by Moss[1], the ultimate detectivity would be $D^*_\lambda = 6\times10^{11}$ cm (c/s)$^{1/2}$ W^{-1}, about 2·5 times greater than the measured value for Te-67-*N*. One disadvantage of this method of estimating D^*_λ is that a knowledge of the intrinsic energy gap is required. In most cases it is somewhat arbitrary as to what value to take as the energy gap. Moss takes the energy gap to be defined by the condition that the photocurrent is 50 per cent of the peak photocurrent. LOFERSKI[2] takes the energy gap to be defined by the condition that the transmission is 5 per cent of its value in the transparent region, i.e. $E_t = 0{\cdot}325$ eV. This corresponds to a $D^*_\lambda = 3\times10^{11}$ cm(c/s)$^{1/2}$ W^{-1}. Another definition sometimes used is the energy corresponding to an absorption coefficient 10^3 times the value in the transmission region. For this energy gap[2], ($E_t = 0{\cdot}275$ eV), $D^*_\lambda = 1{\cdot}6\times10^{11}$ cm (c/s)$^{1/2}$ W^{-1}. Other authors have used still different definitions for the energy gap. For the case of tellurium the definition of the energy gap is further complicated by the fact that the crystal is dichroic. These difficulties can be bypassed to some extent by calculating the ultimate detectivity after the method of Moss[12]. Here a knowledge of the intrinsic energy gap is not necessary. By this method the ultimate detectivity is estimated[14] to be D^*_λ (3·4 μ, —) $= 2{\cdot}2\times10^{11}$ cm (c/s)$^{1/2}$ W^{-1}, compared to the measured value of D^*_λ (3·4 μ, 900) $= 2{\cdot}3\times10^{11}$ cm (c/s)$^{1/2}$ W^{-1}. This excellent agreement between the measured and estimated D^*_λ values is somewhat fortuitous because of the simplifying approximations made in this method of calculation.

From an analysis of these two methods for estimating the theoretical D^*_λ value it is seen that the detectivity for cell Te-67-*N* is probably limited by the fluctuations of the background radiation. A possible explanation as to why the D^*_λ values of the other Te cells are less than that for Te-67-*N* is the relative number of defects in the sensitive cell elements. From the study of whisker growth it is known that the number of defects increases with the crystal diameter. One might then conclude that the Te-*C* elements cleaved from large crystals would intrinsically have a greater number of defects than the Te-*N* elements. Also, as pointed out earlier, tellurium can be easily deformed plastically, and such deformation can introduce a large number of defects. The number of defects must be intimately related to the time constant and thus to the detectivity.[8] Unfortunately, it was not possible to obtain

estimates of the defect densities for any of the Te-cell elements measured once they had been assembled.

REFERENCES

1. T. S. Moss, *Proc. Phys. Soc. Lond.* A**62,** 264 (1949); *Photoconductivity in the Elements.* Butterworth, London (1952).
2. J. J. Loferski, *Phys. Rev.* **93,** 707 (1954).
3. G. H. Suits, Project MICHIGAN Report No. 2144–240–T (1957).
4. C. D. Butter and L. D. McGlauchlin. Presented at the Los Angeles IRIS (Infrared Information Symposium) Meeting, April, 1960.
5. T. J. Davis, *J. Appl. Phys.* **28,** 1217 (1957).
6. J. S. Blakemore, J. D. Heaps, K. C. Nomura and L. P. Beardsley, *Phys. Rev.* **117,** 687 (1960).
7. A. P. de Carvalho, *C.R. Acad. Sci., Paris* **242,** 745 (1956).
8. H. E. Spencer, *J. Appl. Phys.* **31,** 505 (1960).
9. P. Bratt, W. Engeler, H. Levinstein, A. MacRae and J. Pehak, Final Report, Syracuse University, Contract No. AF33 (616)–3859, February, 1960.
10. V. Roberts, Presented at meeting of International Commission for Optics, Stockholm, August, 1959.
11. R. L. Petritz, *Proc. I.R.E.* **47,** 1458 (1959).
12. T. S. Moss, *J. Opt. Soc. Amer.* **40,** 603 (1950).
13. R. A. Smith, F. E. Jones and R. P. Chasmas, *The Detection and Measurement of Infrared Radiation.* Clarendon Press, Oxford (1957).
14. D. F. Edwards and M. Mercado, *Infrared Phys.* **1,** 17 (1961).

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 29–36. Printed in Great Britain

CURRENT DEPENDENCE OF α_{cb} IN THE ALLOY JUNCTION TRANSISTOR

KOGO SATO and MASAMI TOMONO*

Musashi Works, Hitachi Ltd., Kodaira, Tokyo

(*Received* 6 *October* 1960)

Abstract—The emitter-current dependence of α_{cb} of the p–n–p alloy transistor is affected by various factors. One of the important factors at low emitter-current densities is the ratio between donor concentration in base region and acceptor concentration in emitter region. But, at high emitter-current densities these two concentrations are observed experimentally to have essentially different behaviours. When, however, the emitter current is normalized in the form of p/N_d, α_{cb} has a maximum at a point in the vicinity of $p/N_d = 1$ almost irrespective of donor concentration of the base material used. The fall-off of α_{cb} at high emitter current is found to depend on the acceptor concentration in the emitter region. For those transistors with relatively large γ and β, it is ascertained that the approximation of STIELTJES *et al.* agrees fairly well with experimental data. Regarding the effect of surface recombination velocity s on the current dependence of α_{cb}, interesting experimental results are obtained but the quantitative agreement with the approximation theory is rather poor.

Résumé—La façon dont dépend l'α_{cb} du transistor allié p–n–p du courant émetteur est affectée par différents facteurs. Un des facteurs les plus importants à de faibles densités de courant émetteur est le rapport entre la concentration de donneurs dans la région de base et la concentration d'accepteurs dans la région de l'émetteur. Mais à de fortes densités de courant émetteur, on a observé expérimentalement que ces deux concentrations ont de différentes façons de comportement. Quand toutefois le courant émetteur est considéré dans la forme p/N_d, α_{cb} a une valeur maximum à un point dans les environs de $p/N_d = 1$, presqu' indépendamment de la concentration de donneurs dans le matériau de base employé. On a trouvé que le chute d'α_{cb} a un fort courant émetteur dépend de la concentration d'accepteurs dans la région de l'émetteur. Pour les transistors ayant des facteurs γ et β élevés, on s'est assuré que l'approximation de STIELTJES *et al.* est en bon accord avec les données expérimentales. En ce qui concerne l'effet de la vitesse de recombinaison de surface sur la dépendance du courant d'α_{cb}, d'intéressants résultats expérimentaux sont obtenus mais l'accord quantitatif avec la théorie d'approximation est plutôt faible.

Zusammenfassung—Die Emitterstrom-Abhängigkeit des Stromverstärkungsfaktors α_{cb} in einem p–n–p-Transistor mit legiertem Übergang wird durch verschiedene Faktoren beeinflusst. Bei geringer Dichte des Emitterstroms ist einer der wichtigen Faktoren das Verhältnis von Donatorkonzentration in der Basis zu Akzeptorkonzentration im Emitter. Bei hoher Dichte des Emitterstroms zeigen experimentelle Boebachtungen ein ganz verschiedenes Verhalten der beiden Konzentrationen. Wird der Emitterstrom in der Form p/N_d normalisiert, so hat α_{cb} an einem Punkt in der Nähe von $p/N_d = 1$ ein Maximum, und dieses ist von der Donatoremkonzentration des benutzten Basismaterials fast unabhängig. Der Abfall von α_{cb} bei hohem Emitterstrom hängt von der Akzeptorenkonzentration im Emitter ab. Für Transistoren mit relativ hohen γ- und β-Werten stimmt die Annäherung von STIELTJES *et al.* mit den experimentellen Befunden ziemlich gut überein. Im Hinblick auf den Effekt der Rekombinationsgeschwindigkeit s an der Oberfläche auf die Stromabhängigkeit von α_{cb} wurden interessante experimentelle Ergebnisse erzielt, jedoch ist die Übereinstimmung mit der theoretischen Annäherung unbefriedigend.

1. INTRODUCTION

EMITTER-CURRENT dependence of α_{cb} in the alloy junction transistor can be explained by examining it with respect to three factors, surface and bulk recombinations of injected carriers and emitter injection efficiency, as first proposed by

* Central Research Laboratory, Hitachi Ltd., Kokubunji, Tokyo

WEBSTER[1]. Thereafter, RITTNER[2] and others developed his theory, and further FLETCHER and others[3–6] described how injected carriers at high injection level are concentrated near the periphery of the emitter. This causes an increase in the surface recombination velocity and as a consequence the current-amplification factor is decreased. More recently, ARMSTRONG *et al.*[7] and STIELTJES and TUMMERS[8] have reported that the fall-off of α_{cb} at high current densities is remarkably improved by the addition of a small amount of gallium or aluminium to the indium used in the preparation of the emitter. Still further, MOORE and PANKOVE[9], STRIPP and MOORE[10], MORTENSON[11] and YANAI and SUGANO[12] have described the influence of transistor geometry on the current dependence of α_{cb}. This article presents the results of an experimental study of the effects of donor concentration in the base region, acceptor concentration in the emitter region, surface recombination velocity and perfection of the junction on the emitter-current dependence of α_{cb} in a practical current-density region. Also given are results of an examination of the applicability of the STIELTJES and other theories to actual transistors. In our experiment, p–n–p transistors with planar junctions were used with the view of avoiding the influence of geometry, as far as possible.

2. METHOD OF EXPERIMENT

The germanium used in the experiment was an n-type single crystal doped with Sb. It had a lifetime exceeding 300 μsec and an etch-pit density of 3000–6000 cm^{-2}. A pure indium dot 1·26 mm in diameter was used as the collector. A pure indium dot or In–Ga–Ag alloy dot 0·8 mm in diameter was used as the emitter. They were alloyed in hydrogen gas at 560°C for 5 min. Our experiments were made using transistors with planar p–n junctions, as shown in the photomicrograph of Fig. 1.

The surface recombination velocity for a transistor s is difficult to measure, and, instead, the effective lifetime of minority carriers, τ_{eff}, was measured by the method of LEDERHANDLER–GIACOLETTO[13] using electropulse techniques. The measurement of the grounded-emitter, small-signal current-amplification factor α_{cb} was made in all cases at 6-V constant collector voltage and with the transistor case dipped into a mercury bath to minimize the effect of heat generated by the power dissipation inside the transistor.

3. RESULTS OF EXPERIMENT

1. *Effect of donor concentration in the base region*

Fig. 2 shows the effect of donor concentration

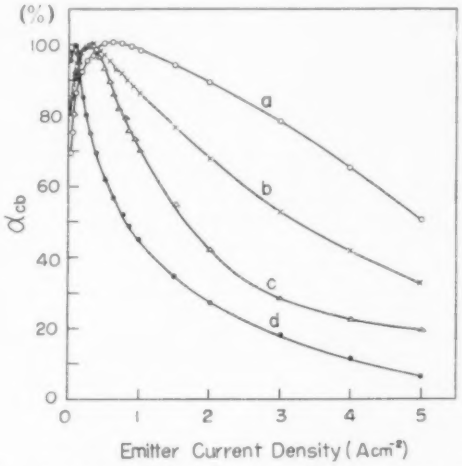


FIG. 2. Effect of the donor concentration in the base on the current dependence of α_{cb} (pure In emitter). $N_a \simeq 5\times10^{18}$ cm^{-3}. (a) N_d $1{\cdot}7\times10^{14}$ cm^{-3}; (b) N_d $1{\cdot}7\times10^{14}$ cm^{-3}; (c) N_d $1{\cdot}2\times10^{14}$ cm^{-3}; (d) N_d $4{\cdot}3\times10^{13}$ cm^{-3}.

in the base region on the emitter-current dependence of α_{cb} of the p–n–p transistor which was made by alloying indium to an n-type germanium pellet and had nearly constant effective lifetime τ_{eff} of a hole.

Curves (a–d) in the figure are plotted respectively for four values of donor concentration in the base region varying from $1{\cdot}7\times10^{15}$ cm^{-3} to $4{\cdot}3\times10^{13}$ cm^{-3}. It is seen that as the donor concentration in the base region is decreased, the rate of fall-off of α_{cb} becomes appreciable and the emitter current at which α_{cb} reaches a maximum moves toward the lower current-density side.

Fig. 3 shows the current dependence of α_{cb} which was obtained rather statistically from each of thirty samples with various donor concentrations. Similarly to the case of Fig. 2, it indicates that the current dependence of α_{cb} becomes greater with decreasing impurity concentration in the base region.

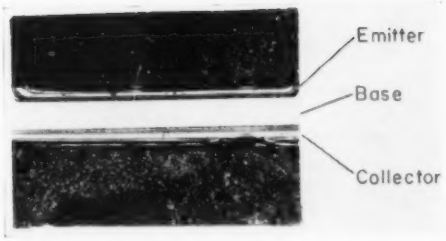


FIG. 1. Cross-section of the transistor used in experiment.

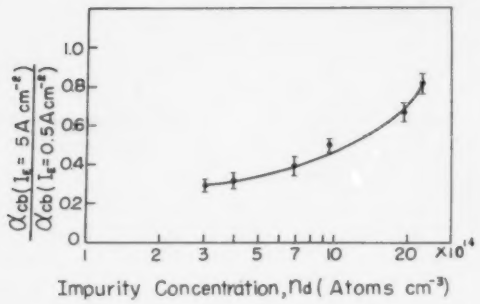


FIG. 3. Effect of donor concentration in the base on the current dependence of α_{cb}.

The ordinate of this figure represents the ratio of α_{cb} at an emitter-current density of 0·5 A-cm^{-2} to that of 5 A-cm^{-2} and serves as a practical measure showing the current dependence of α_{cb}.

2. *Effect of acceptor concentration in the emitter region*

Fig. 4 shows the relationship between equilibrium donor concentration in the base region N_d and emitter-current dependence of α_{cb}, with the value of N_d varying from $1{\cdot}7\times10^{15}$ cm^{-3} to $5{\cdot}6\times10^{13}$ cm^{-3}, as in experiment (1).

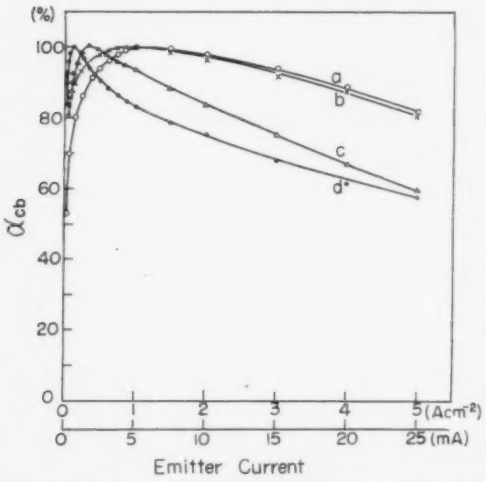


FIG. 4. Effect of the donor concentration in the base region on the current dependence of α_{cb} (In–Ga–Ag emitter). $N_a \simeq 5\times10^{19}$ cm^{-3}. (a) N_d $1{\cdot}7\times10^{15}$ cm^{-3}; (b) N_d $1{\cdot}7\times10^{14}$ cm^{-3}; (c) N_d $1{\cdot}2\times10^{14}$ cm^{-3}; (d) N_d $5{\cdot}6\times10^{13}$ cm^{-3}.

The emitter metal used was prepared by adding a small amount of gallium to the indium to increase the equilibrium acceptor concentration N_a in the emitter region. As clearly seen from the figure, compared with the pure indium case of Fig. 2 the fall-off of α_{cb} is very small and the emitter current at which α_{cb} is a maximum moves relatively to the high-current-density side. That is, in the former the maximum point appears in the range of 0·1 to 0·7 A-cm^{-2}, while in the latter it appears in the range of 0·2 to 1·2 A-cm^{-2}. With respect to the donor concentration in the base region of $1{\cdot}7\times10^{15}$ cm^{-3}, α_{cb} decreases to about 50 per cent at the current density of 5 A-cm^{-2} in the case of the pure indium emitter but only to 82 per cent in the case of the In–Ga–Ag emitter. As for the current density at which α_{cb} is a maximum, it is 0·6 A-cm^{-2} for the former but 1·2 A-cm^{-2} for the latter.

3. *Effect of surface-recombination velocity*

Because of reasons to be described later, surface-recombination velocity of holes and electrons also has a marked effect on the current dependence of α_{cb}. The effect is especially noticeable at relatively low current densities. Our experiment indicates that for those alloy-junction transistors with the same geometry and emitter efficiency γ,

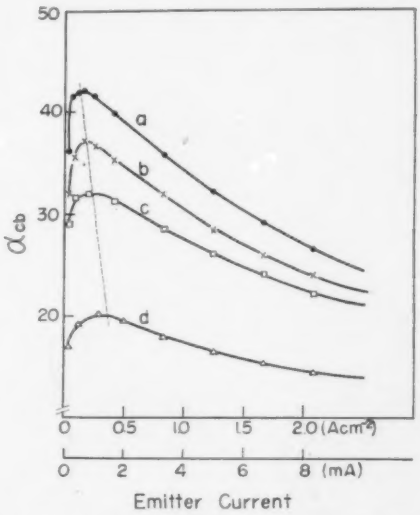


FIG. 5. Effect of the effective lifetime τ_{ett} of minority carrier on the current dependence of α_{cb}. (a) τ_{eff} $8{\cdot}33\times10^{-6}$ sec; (b) τ_{eff} $7{\cdot}95\times10^{-6}$ sec; (c) τ_{eff} $6{\cdot}70\times10^{-6}$ sec; (d) τ_{ett} $4{\cdot}87\times10^{-6}$ sec.

α_{cb} in the vicinity of 0·25 A-cm^{-2} is determined uniquely by τ_{eff} with the empirical relation $\alpha_{cb} \simeq (268\,\tau_{eff})/(34{\cdot}8-\tau_{eff})$. At high current densities, however, the situation is too complicated to be expressed by such a simple expression, since the current dependence of emitter injection efficiency γ should simultaneously be taken into account, as mentioned later.

In order to examine the effect of surface-recombination velocity on the current dependence of α_{cb}, a single transistor element was surface-treated in various ways, and at each time the $\alpha_{cb}-I_E$ characteristic was measured. The results of these experiments are shown in Fig. 5. As seen from this figure, the effect of surface-recombination velocity s is distinctive at relatively low current densities but is trivial at high current densities. Moreover, as τ_{eff} becomes small, that is as s becomes large, the current density at which α_{cb} reaches a maximum moves toward the high-current-density side. This is the result which is also expected from the discussion appearing later. With the aim of varying the value of s in a wider range, a single transistor element was dipped into organic solvents with different dielectric constants. With τ_{eff} varied in the range of 3×10^{-6} sec to $1{\cdot}5\times10^{-5}$ sec in this way, the $\alpha_{cb}-I_E$ characteristic was extensively changed as seen from Fig. 6.

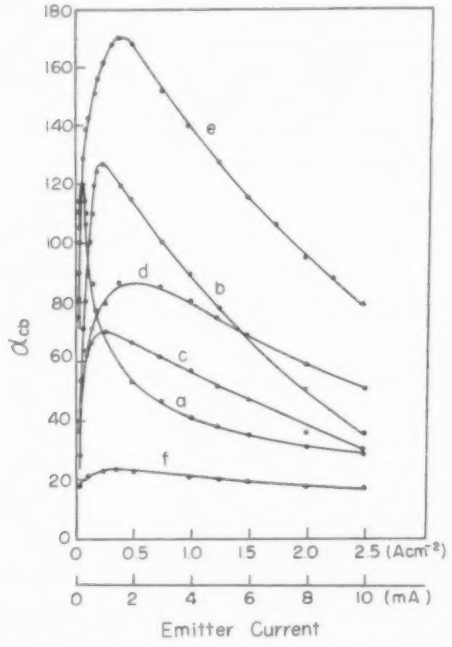


FIG. 6. Effect of the organic solvents on the current dependenc eof α_{cb}. (a) Dry air (τ_{eff} 12 μsec); (b) nitrobenzene (τ_{eff} 9·2 μsec); (c) acetone (τ_{eff} 8·7 μsec); (d) aniline (τ_{eff} 8·2 μsec); (e) alcohol (τ_{eff} 5·1 μsec); (f) vacuum bake (τ_{eff} 3·6 μsec).

In Fig. 6, with the exceptions of curves (e) and (f), the smaller the value of τ_{eff} the more the maximum point of α_{cb} moves to the high-current-density side and the smaller the fall-off factor becomes. But the relative values of α_{cb} and the current dependence cannot be accounted for by the effect of τ_{eff} alone.

4. *Effect of other factors*

Fig. 7 shows the relationship between the alloying temperature and the current dependence of

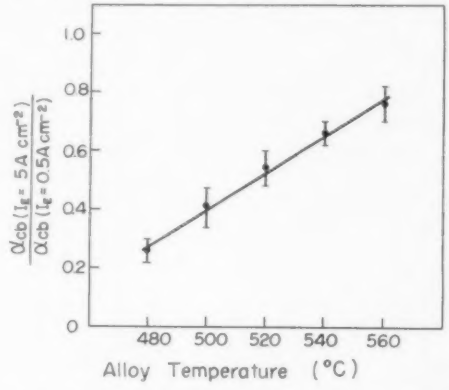


FIG. 7. Relationship between the alloying temperature and the current dependence of α_{cb}.

α_{cb}. The higher the alloying temperature the smaller the current dependence of α_{cb}. It has been ascertained by various other methods that the perfectness of the p–n junction is closely related to the alloying temperature. And in the present work also, it is conceivable that the higher the alloying temperature the fewer the number of weakpoints in the p–n junction, the γ is improved and consequently the current dependence of α_{cb} is improved. The values of γ at low current density are about 0·941 and 0·999 for transistors which were alloyed at 480°C and 560°C, respectively.

In addition, the base width W_b should have some influence on the current dependence of α_{cb}.

Fig. 8, for instance, shows the current dependence of α_{cb} for the base widths of 15×10^{-4} cm and 25×10^{-4} cm, with the other conditions held nearly constant.

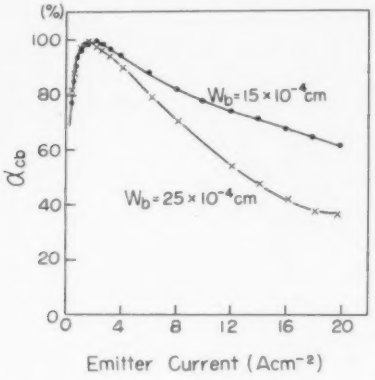


FIG. 8. Effect of the basewidth W_b on the current dependence of α_{cb}.

From this figure it is obvious that the fall-off is less for the former and that the base width is one of the factors which must be taken into consideration in designing transistors to be used at high current densities.

4. DISCUSSION

In this section, first the general treatments by WEBSTER[1], STIELTJES and TUMMERS[8] and others of the current dependance of α_{cb} are briefly reviewed and then their correspondence with our experimental results is examined.

1. α_{cb} *for low injection level*

The current-amplification factor of the alloy-junction transistor α_{cb} is expressed in a general form as shown by equation (1). In the case that injection of holes into the base region is so small ($1\gg p/N_d$) that the resulting reduction in γ and influence on the static field may be neglected, it can be reduced to the form of equation (2).

$$\frac{1}{\alpha_{cb}}=\frac{\partial I_{SR}}{\partial I_{Ep}}+\frac{\partial I_{VR}}{\partial I_{Ep}}+\frac{\partial I_{Ee}}{\partial I_{Ep}} \tag{1}$$

$$\frac{1}{\alpha_{cb}}=\frac{sA_sW_b}{D_pA_e}+\frac{1}{2}\left(\frac{W_b}{L_p}\right)^2+\frac{\sigma_eW_b}{\sigma_eL_e} \tag{2}$$

where s = surface-recombination velocity, A_s = effective area for surface recombination, W_b = base width, D_p = diffusion coefficient for holes in the base region, A_e = area of emitter, σ_b, σ_e = conductivities in base and emitter regions and L_e, L_p = diffusion length for electrons in the emitter region and for holes in the base region.

2. α_{cb} *for high injection level*

(i) *Static field caused by the injected carrier in base region.* Hole-current density J_p and electron-current density J_e of the base region are given, if the static field is ignored, respectively by

$$J_p=-q\,D_p\,\text{grad}\,p \tag{3}$$

$$J_e=q\,D_n\,\text{grad}\,n \tag{4}$$

where p and n denote the hole and electron concentrations in the base region, respectively. If the static field is taken into account, J_p and J_e are given as

$$J_p=-p\,q\,\mu_p\,\text{grad}\,V-q\,D_p\,\text{grad}\,p \tag{5}$$

$$J_e=-n\,q\,\mu_n\,\text{grad}\,V+q\,D_n\,\text{grad}\,n \tag{6}$$

By taking into account the neutrality condition valid for the base region and using equations (5) and (6), J_p is expressed as

$$J_p=-q\,D_p\,\text{grad}\,p\left(1+\frac{p}{N_d}\right) \tag{7}$$

where N_d = equilibrium donor concentration in the base region and p = injected hole concentration in the base region.

If the injected hole concentration p is very large compared with the equilibrium donor concentration N_d in the base region, the hole-current density J_p is doubled:

$$J_p\simeq-2\,q\,D_p\,\text{grad}\,p \tag{8}$$

From this equation, the variation of D_p with respect to the injection level (p/N_d) can be graphically represented as shown in Fig. 9.

In this figure, D_p' is an effective diffusion coefficient due to the static field caused by the injected hole in the base region, and is expressed by the relation $D_p'=D_p\,[1+p/(N_d+p)]$. With increasing injected-hole concentration, p and D_p'/D_p gradually approach 2.

(ii) *Corrections to current dependence of* α_{cb}. According to WEBSTER, as a compensation for the surface-recombination velocity s and emitter-

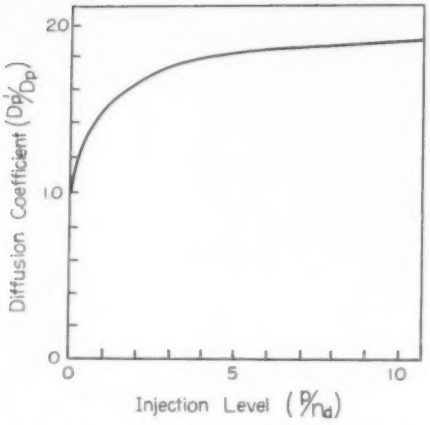


FIG. 9. Relationship between the injection level and the effective diffusion coefficient.

injection efficiency γ at high injection level, a quantity called the "field factor" $g(Z)$ is introduced. The expression as such, however, is not favourable for separating the effect of the earlier-mentioned donor concentration in the base region and acceptor concentration in the emitter region on the $\alpha_{cb}-I_E$ characteristic. It is more convenient to follow the procedure of STIELTJES and others. Firstly, the surface-recombination velocity affects the base efficiency β. The effect is expressed, as seen from the aforementioned discussion, by equation (9) which takes into account the current dependence of the hole diffusion coefficient D_p

$$(1-\beta) = \frac{W_b^2}{2\tau_{eff}D_p}\left(\frac{1+p/N_d}{1+2p/N_d}\right) \qquad (9)$$

On the other hand, the current dependence of the emitter-injection efficiency is given by

$$(1-\gamma) = \frac{D_n\bar{W}_bN_d}{D_pL_eN_a}\left(1+\frac{p}{N_d}\right) \qquad (10)$$

It follows from the above that $(1-\alpha)$ at each injection level can be approximated as follows:

$$(1-\alpha) \simeq (1-\beta)+(1-\gamma)$$
$$\simeq \frac{W_b^2}{2\tau_{eff}D_p'}+\frac{D_n\bar{W}_bN_d}{D_pL_eN_a}\left(1+\frac{p}{N_d}\right) \qquad (11)$$

As is easily understood, the first and second terms in equation (11) contribute as the rise factor and fall-off factor of α_{cb}, respectively.

3. *Application to the experimental results*

(i) *Maximum condition for* $\alpha_{cb}-I_E$ *curve.* The measured emitter currents of Figs. 2 and 4 presented in the earlier section are shown in Fig. 10 and Fig. 11 respectively as re-expressed in terms of the injection level p/N_d.

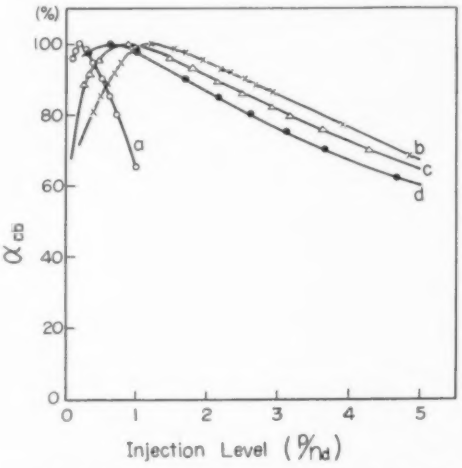


FIG. 10. Relationship between the injection level and α_{cb} (pure In emitter). $N_a \simeq 5\times10^{18}$ cm^{-3}. (a) $N_d = 1{\cdot}7\times10^{15}$ cm^{-3}; (b) $N_d = 1{\cdot}7\times10^{14}$ cm^{-3}; (c) $N_d = 1{\cdot}2\times10^{14}$ cm^{-3}; (d) $N_d = 4{\cdot}3\times10^{13}$ cm^{-3}.

It is obvious from these figures that except for curve (a) corresponding to the donor concentration in the base region of $N_d = 1{\cdot}7\times10^{15}$ cm^{-3}, the fall-off of α_{cb} for each curve in both figures is relatively independent of the donor concentration in the base region and that the points at which α_{cb} reaches a maximum appear in the vicinity of $p/N_d = 1$. These points, however, can be moved somewhat arbitrarily by properly selecting the value of s, as shown in Fig. 5. The appearance of maxima in the vicinity of $p/N_d = 1$ may be attributed to the fact that the effect of reduction in γ becomes distinctive at hole injection levels of $p/N_d = 1$ as expected from equation (11).

(ii) *Current dependence of* $(1-\gamma)$. The next subject to be examined is the actual validity of the approximate equation (11). Fig. 12 shows the

comparison of calculated and measured effects of γ on the current dependence of α_{cb} with respect to those samples with different values of N_d, W_b, etc., as shown in Table 1, that is, the different values of γ.

It is seen from this figure that for samples No. 10 and No. 15 having $(1-\gamma)$ below a certain value, the values calculated from equation (11) agree well with the measured values. But, for sample No. 1, having large $(1-\gamma)$, the calculation does not agree well with experiment. This indicates that equation (11) is a relatively good approximate formula for transistors with small $(1-\gamma)$.

(iii) *Current dependence of* $(1-\beta)$. Fig. 13 shows the same current dependence of α_{cb} as given in Fig. 5 with $(1-\beta)$ varied by the value of τ_{eff} but

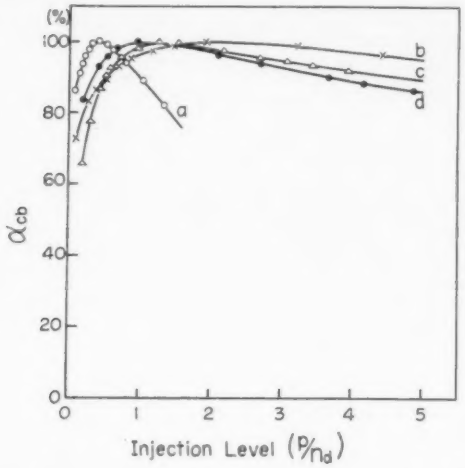


FIG. 11. Relationship between the injection level and α_{cb} (In–Ga–Ag emitter). $N_a \simeq 5\times10^{19}$ cm^{-3}. (a) $N_d = 1{\cdot}7\times10^{15}$ cm^{-3}; (b) $N_d = 1{\cdot}7\times10^{14}$ cm^{-3}; (c) $N_d = 1{\cdot}2\times10^{14}$ cm^{-3}; (d) $N_d = 5{\cdot}6\times10^{13}$ cm^{-3}.

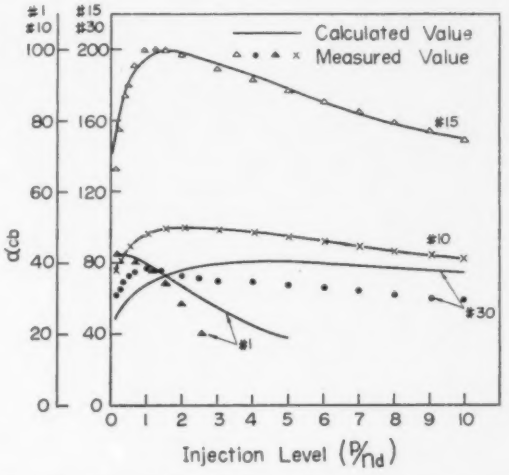


FIG. 12. Comparison of theory and experiment. (In–Ga–Ag emitters, No. 1, $N_d = 1{\cdot}7\times10^{15}$ cm^{-3}; No. 10, $N_d = 1{\cdot}7\times10^{14}$ cm^{-3}; No. 15, $N_d = 1{\cdot}2\times10^{14}$ cm^{-3} and No. 30, $N_d = 4\times10^{13}$ cm^{-3}.)

Table 1

Parameter		Sample No.			
		1	10	15	30
Emitter area (10^{-3}cm^2)	A_e	4·1	4·3	4·3	4·2
Acceptor concentration in emitter region (10^{19}cm^{-3})	N_a	5·0	5·0	5·0	5·0
Donor concentration in base region (10^{14}cm^{-3})	N_d	17	1·7	1·2	0·4
Base width (10^{-3}cm)	W_b	4·0	5·0	2·5	6·0
Effective lifetime (10^{-5} sec)	τ_{eff}	1·0	1·0	1·2	1·1
Diffusion coefficient for holes (cm^{2}sec^{-1})	D_p	44	44	44	4·4
Diffusion coefficient for electrons (cm^2 sec^{-1})	D_n	92	92	92	92
Diffusion length for electrons in emitter (10^{-5}cm)	L_e	4·4	4·4	4·4	4·4

with $(1-\gamma)$ held constant, in comparison with curves (a), (c) and (d) showing the values calculated with the use of equation (11) with respect to the same values of τ_{eff}. As is clearly seen from this figure, curve (a) for relatively large τ_{eff} agrees relatively well with the measurement. But curves (c) and (d) for small τ_{eff} differ considerably from the measurement, especially in the case of (d).

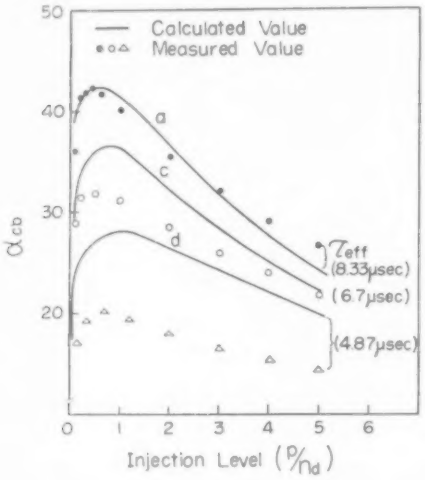


FIG. 13. Effect of τ_{eff} on the current dependence of α_{cb} (curves show the calculated values and ●, ○, △ the measured ones).

This suggests that the use of τ_{eff} measured at low current density (0·13 A-cm^{-2}) is improper for high current densities and that the approximate formula (11) is not valid for transistors with large $(1-\beta)$.

In respect of this, and also to understand the interesting phenomenon of Fig. 6, it will be necessary in the future to measure the value of τ_{eff} at each injection level.

Regarding the fall-off of α_{cb} in a region with lower current densities than the current region which has been dealt with here, it seems necessary to introduce the effect of carrier generation and recombination inside the depletion layer of the p–n junction and in other places, as proposed by GÄRTNER and HALL[14].

Acknowledgements—The authors wish to express their appreciation and indebtedness to S. TAUCHI for his supply of the single crystal, to T. TOKUYAMA for his co-operation in the measurement of τ_{eff}, and to T. SUZUOKA, M. OZAWA and K. KOBAYASHI for their co-operation in the experiments.

REFERENCES

1. W. M. WEBSTER, *Proc. I.R.E.* **42,** 914 (1954).
2. E. S. RITTNER, *Phys. Rev.* **94,** 1161 (1954).
3. R. U. PRITCHARD and W. N. COFFER, *I.R.E. Convention Record, Wisconsin*, Vol. 2, p. 89 (1954).
4. N. H. FLETCHER, *Proc. I.R.E.* **43,** 551 (1955).
5. R. EMEIS and A. HERLET, *Proc. I.R.E.* **46,** 1216 (1958).
6. S. WANG and T. T. WU, *I.R.E. Trans.* **Ed-6,** 162 (1959).
7. L. D. ARMSTRONG, C. L. CARLSON and M. BENTIVEGNA, *RCA Rev.* **17,** 37 (1956).
8. F. H. STIELTJES and L. J. TUMMERS, *Philips Tech. Rev.* **18,** 61 (1956).
9. A. R. MOORE and J. I. PANKOVE, *Proc. I.R.E.* **42,** 907 (1954).
10. K. F. STRIPP and A. R. MOORE, *Proc. I.R.E.* **43,** 856 (1955).
11. K. E. MORTENSON, *I.R.E. Trans.* **Ed-6,** 174 (1959).
12. H. YANAI and T. SUGANO, *Proc. I.R.E.* **48,** 1739 (1960).
13. S. R. LEDERHANDLER and L. J. GIACOLETTO, *Proc. I.R.E.* **43,** 447 (1955).
14. W. W. GÄRTNER and R. HALL, *Proc. I.R.E.* **46,** 1875 (1958).

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 37–44. Printed in Great Britain

CONSTITUTIONAL SUPERCOOLING DURING CRYSTAL GROWTH FROM STIRRED MELTS—I

THEORETICAL

D. T. J. HURLE

Ministry of Aviation, Royal Radar Establishment, Great Malvern, Worcestershire

(*Received* 29 *July* 1960; *in revised form* 30 *January* 1961)

Abstract—The theory of constitutional supercooling is applied to the case of crystal growth from a stirred melt. Expressions, applicable to crystal growth using the Czochralski technique, are derived which predict the onset of constitutional supercooling and the magnitude of the supercooling when it exists.

Résumé—La théorie de surfusion de constitution est appliquée au cas du développement de cristal, en partant de la fonte remuée. Des formules applicables au développement du cristal employant la technique de Czochralski sont derivées; elles prédisent le début de la surfusion de constitution et la qualité de la surfusion quand elle existe.

Zusammenfassung—Die Theorie der Unterkühlung wird auf das Kristallwachstum aus einer gerührten Schmelze angewendet. Für das Kristallwachstum nach dem Czochralski-Verfahren werden Ausdrücke abgeleitet, aus denen sich das Einsetzen und das Ausmass der Unterkühlung vorausbestimmen lassen.

INTRODUCTION

SUPERCOOLING resulting from the redistribution of solute at the solid–liquid interface of a growing metal crystal (so-called "constitutional supercooling") was pointed out by RUTTER and CHALMERS[1], and the theory has been worked out by TILLER *et al.*[2] for the case of an unstirred melt. The validity of the theory has been experimentally tested, using horizontal crystal growth techniques, by TILLER and RUTTER[3], for lead with known additions of tin, silver and gold; by WALTON *et al.*[4] for tin with lead; and by PLASKETT and WINEGARD[5] for tin with bismuth, antimony and lead.

Constitutional supercooling is not, in general, encountered during the growth of semiconductor crystals at the usual levels of doping. However, at the high levels of doping required for the production of material for ESAKI diodes,[6] constitutional supercooling of the melt can occur.

Since it is common practice to grow single crystals of the widely used semiconductor materials by the Czochralski method in which the melt is stirred by rotating the crystal as it is "pulled", it would appear useful to extend the theory of constitutional supercooling to stirred melts. In this paper expressions are derived for the onset of constitutional supercooling and the magnitude of the supercooling, when it exists, in a stirred melt.

THEORY

When growth proceeds into an impure melt containing a solute whose distribution coefficient (k) is less than unity, solute is continually being rejected into the melt at the interface, resulting in a layer containing an enhanced solute concentration immediately ahead of the interface. In the absence of stirring in the melt, solute is removed from the layer only by diffusion down the concentration gradient into the bulk melt. TILLER *et al.*[2] have shown that a steady state is reached when the solute has "piled up" to a concentration $C_0 = C_L/k$ at the interface. If the melt is stirred, however, then C_0 will have some lower value lying between C_L/k and C_L. C_L is the solute concentration in the bulk melt. The region where the concentration at the interface is building up from C_L to

its steady-state value is called the transient region.

In considering the solute distribution in a melt which is stirred by rotating the crystal, BURTON and

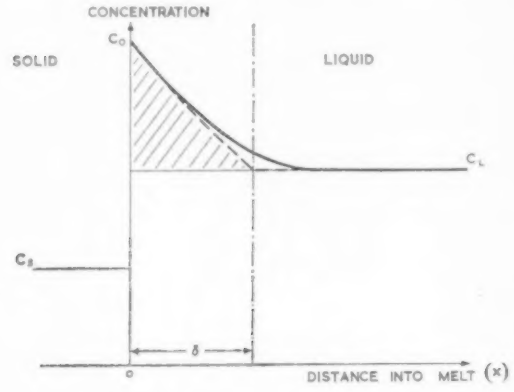


FIG. 1. Solute distribution in the melt ahead of the solid–liquid interface. (—— Actual solute distribution; - - - - mathematical approximation to solute distribution.)

SLICHTER[7] and BURTON *et al.*[8] divided the melt into two regions (Fig. 1). The first, of width δ, in which solute transport was supposed to be by diffusion only, and the second, the bulk melt, in which the stirring was supposed to ensure complete mixing of the solute in the solvent. If the actual solute distribution in the melt is as shown in Fig. 1, then the value of δ is chosen so that the solution to the diffusion equation applied for $0 \leqslant x \leqslant \delta$ fits the actual distribution accurately at $x = 0$. The other principal assumptions in their theory are as follows.

(a) Thermal and solute flow is one-dimensional in the direction of growth (x-direction).

(b) Solute diffusion in the solid is negligible.

(c) The interface is a plane disk normal to the growth direction.

(d) The distribution coefficient is independent of concentration.

Assuming an infinite melt, the concentration in the liquid at the interface will build up to the steady value C_0, whereafter $dC/dt = 0$. The solute diffusion equation can then be written, with respect to a co-ordinate system having its origin at the interface, as

$$D\frac{d^2C}{dx^2} - V_x\frac{dC}{dx} = 0 \qquad (1)$$

where D is the diffusion coefficient of the solute in the solvent and V_x is the flow velocity in the x-direction. By taking $V_x = -f$ (where f is the growth rate), i.e. by ignoring the contribution of the fluid-flow velocity to V_x, BURTON and SLICHTER[7] obtain the following solution to equation (1).

$$\left(\frac{dC}{dx}\right) = \frac{f}{D}(C_s - C_0)\exp(-X) \qquad (2)$$

and

$$\frac{C_x - C_s}{C_L - C_s} = \exp(\Delta - X) \qquad (3)$$

where $\Delta = f\delta/D$ and $X = fx/D$, and C_s is the concentration in the solid. BURTON and SLICHTER[7] showed that equations (2) and (3) accurately represent the solute distribution at $x = 0$ if the value of δ is given by

$$\delta = 1{\cdot}6\, D^{1/3}\nu^{1/6}\omega^{-1/2} \qquad (4)$$

where ω is the angular velocity of the crystal and ν is the kinematic viscosity of the melt. The equations (2) and (3) are meant to apply to the region $0 \leqslant x \leqslant \delta$, but they become progressively more inaccurate as $x \to \delta$. The limitation $\Delta \leqslant 1$ is imposed on the equations.

The equilibrium liquidus temperature (T_x^E) at any point x in the melt ($0 \leqslant x \leqslant \delta$) is

$$T_x^E = T^0 + mC_x \qquad (5)$$

where m is the slope of the liquidus line ($m < 0$ for $k < 1$) and T^0 is the melting temperature of the pure solvent. [m is approximately constant at low concentrations for an ideal solution; see equation (25).] The equilibrium liquidus temperature gradient at the interface from (2) and (5) is therefore

$$\left(\frac{dT^E}{dx}\right)_{x=0} = \frac{mf}{D}(C_s - C_0) \qquad (6)$$

To obtain equation (6) in terms of C_L, we use the expression of BURTON *et al.*[8] for the effective distribution coefficient (k_{eff})

$$k_{\text{eff}} = \frac{k}{k + (1-k)\exp(-\Delta)} \qquad (7)$$

where

$$k_{\text{eff}} = C_s/C_L \text{ and } k = C_s/C_0$$

Equation (6) becomes

$$\left(\frac{dT^E}{dx}\right)_{x=0} = -\frac{mfC_L(1-k)}{D\{k+(1-k)\exp(-\Delta)\}} \quad (8)$$

The actual temperature gradient in the melt immediately ahead of the interface (G_L) is determined by the condition of conservation of heat flux at the interface, viz.

$$K_L G_L = K_s G_s - L\rho f \quad (9)$$

where G_s = temperature gradient in the solid at the interface, L = latent heat of fusion of the solvent and ρ = density; K_s, K_L = thermal conductivity of the solid and liquid phases respectively. (The volume change on solidification is ignored.)

BURTON and SLICHTER (p. 121 *et seq.*)[7] have examined in detail the temperature gradient in the melt ahead of the interface and conclude that G_L is constant at least out to a distance δ with a value given by equation (9).

Constitutional supercooling will be avoided if the actual temperature gradient in the melt is greater than, or equal to, the equilibrium liquidus temperature gradient at the interface: i.e. if, from (8) and (9),

$$\frac{K_s G_s - L\rho f}{K_L} \geqslant -\frac{mfC_L(1-k)}{D\{k+(1-k)\exp(-\Delta)\}} \quad (10)$$

Fig. 2 shows the equilibrium liquidus temperature distribution in the melt (solid line) and also three possible "real" temperature distributions in the melt (dashed lines). Constitutional supercooling is just avoided with line (1) but is present for lines (2) and (3). The effect of the solute distribution is to give a gradient of supercooling (dS/dx) away from the interface. The magnitude of the effect can be expressed by $(dS/dx)_{x=0}$ which is

$$\left(\frac{dS}{dx}\right)_{x=0} = -\frac{mfC_L(1-k)}{D\{k+(1-k)\exp(-\Delta)\}} - \frac{K_s G_s - L\rho f}{K_L} \quad (11)$$

The magnitude of $(dS/dx)_{x=0}$ increases with increasing growth rate (f) and decreasing stirring rate (ω) at a given melt concentration (C_L). However, expressed in terms of C_s, equation (11) becomes

$$\left(\frac{dS}{dx}\right)_{x=0} = -\frac{mfC_s}{D}\,\frac{(1-k)}{k} - \frac{K_s G_s - L\rho f}{K_L} \quad (12)$$

which is independent of ω for the incorporation of a given concentration of solute (C_s) in the solid

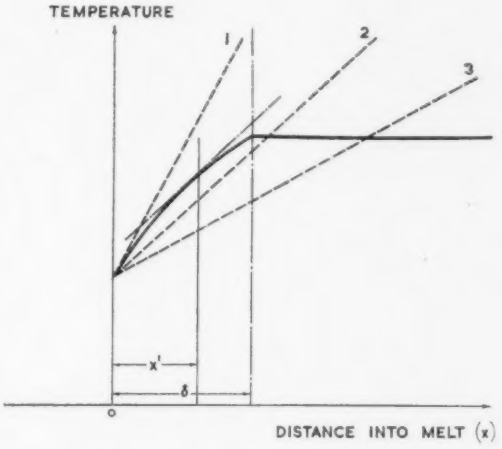


FIG. 2. Temperature distribution in the melt ahead of the solid–liquid interface. (——— Equilibrium liquidus temperature; – – – – actual temperature of melt.)

(provided $\Delta \ll 1$). It will be noted that solutes having small values of k are the most effective in producing constitutional supercooling. In Fig. 3, values of C_s are plotted against f for which $(dS/dx)_{x=0} = 0$ for the case of germanium containing solutes having differing values of the parameter $p = -m(1-k)/Dk$. The data used are

$K_s = 0{\cdot}25$ W/°C-cm; $K_L = 0{\cdot}47$ W/°C-cm
$L = 114{\cdot}3$ cal/g; $\rho = 5{\cdot}4$ g/cm^3
$G_s = 100$°C/cm

C_s varies as $1/f$ for small f because $L\rho f \ll K_s G_s$. For large values of f, C_s decreases rapidly and becomes zero at a growth rate (f_0) at which the temperature gradient in the melt (G_L) is zero.

$$f_0 = \frac{K_s G_s}{L\rho}$$

($= 9{\cdot}7 \times 10^{-3}$ cm/sec for the case considered).

The supercooling (S_x) at any point $x\,(0 \leqslant x \leqslant \delta)$, when it exists, is

$$S_x = (T_0 + mC_x) - (T_0 + mC_0 + G_L x) \\ = -m(C_0 - C_x) - G_L x \tag{13}$$

The magnitude of the supercooling will be greatest at some point $x = x'$ at which $(dS/dx)_{x=x'} = 0$ (see Fig. 2). The value of x' can be obtained by differentiating equation (13) with respect to x and equating to zero at $x = x'$. Thus

$$m\left(\frac{dC}{dx}\right)_{x=x'} = G_L \tag{14}$$

C_x can be obtained by combining equations (3) and (7), but the accuracy of the value becomes progressively poorer as x/δ approaches unity (Fig. 1). To obtain C_x accurately it is necessary to solve equation (1) taking into account the fluid velocity due to the crystal rotation (W), i.e. by writing $-V_x = f + W$ (assuming that the velocities due to growth and fluid flow are additive). For the fluid velocity normal to a rotating disk, COCHRAN[9] obtains

$$W = -(\nu\omega)^{1/2}[a\zeta^2 - \tfrac{1}{3}\zeta^3 - \tfrac{1}{6}b\zeta^4 \ldots .] \tag{15}$$

where $a = 0{\cdot}510$, $b = -0{\cdot}616$ and $\zeta = (\nu/\omega)^{-1/2}x$.

Equation (15) can be approximated to $W \sim 0{\cdot}510\ \omega^{3/2}\ \nu^{-1/2}\ x^2$ for $x < (\nu/\omega)^{1/2}$. (When $x = (\nu/\omega)^{1/2}$, the second term in the expansion is 0·65 times the first term.)

With the boundary condition

$$(C_0 - C_s)f + D\frac{dC}{dx} = 0 \quad \text{at } x = 0$$

equation (1) integrates to

$$\frac{dC}{dx} = -\frac{f}{D}(C_0 - C_s)\exp(-X - BX^3) \tag{16}$$

where

$$B = \frac{0{\cdot}510\ \omega^{3/2}\ D^2}{3\nu^{1/2}f^3}$$

Combining equations (14) and (16)

$$pfC_s \exp(-X' - BX'^3) = G_L \tag{17}$$

where $X' = fx'/D$.

Using Cardan's formula we find that (17) has the following real root

$$X' = \left\{-\frac{C}{2B} + \left(\frac{C^2}{4B^2} + \frac{1}{27B^3}\right)^{1/2}\right\}^{1/3} \\ + \left\{-\frac{C}{2B} - \left(\frac{C^2}{4B^2} + \frac{1}{27B^3}\right)^{1/2}\right\}^{1/3} \tag{18}$$

where $C = \ln(G_L/pfC_s)$.

Using the limitation that $x < (\nu/\omega)^{1/2}$, then

$$x/\delta < 0{\cdot}63\ (\nu/D)^{1/3}$$

Substituting typical values for ν and D for germanium alloys: viz.
$\nu = 2{\cdot}5 \times 10^{-3}$ cm^2 sec^{-1} and $D = 10^{-4}$ cm^2 sec^{-1}

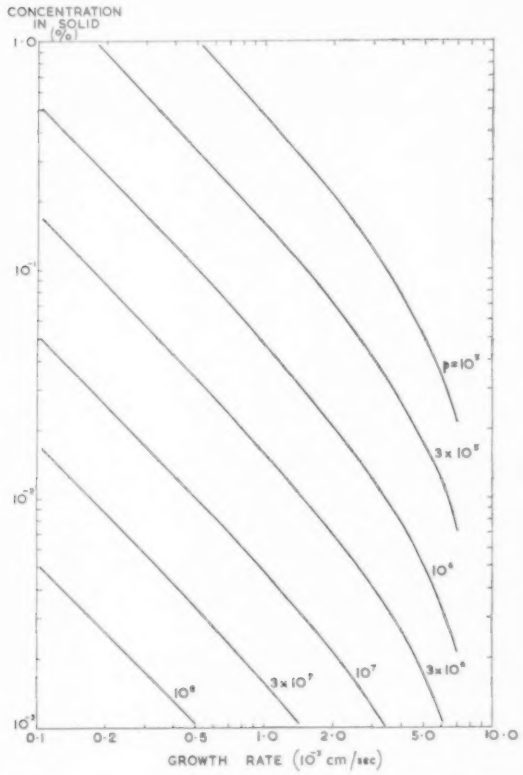


FIG. 3. Concentration of solute in the solid vs. growth rate for which $(dS/dx)_{x=0} = 0$ for germanium containing solute elements having different values of the parameter $p = -m(1-k)/Dk$.

gives

$$x/\delta = X/\Delta < 1{\cdot}84$$

We are therefore restricted to computing the dependence of C_x on x in the region $0 \leqslant x < 1{\cdot}84\delta$.

BURTON and SLICHTER[7] integrated equation (16) numerically and fitted it to equation (3) to obtain their expression for δ. They found that Δ was given by $B\,\Delta^3 = A^3$ where A varied slightly with Δ, being: $A = 0{\cdot}890 - 0{\cdot}060\Delta$ for $-1 < \Delta < +1$. Further since we are restricted to $\Delta \leqslant 1$ and $X < 1{\cdot}84\Delta$, we see that the maximum numerical value which the exponent on the right-hand side of equation (16) can take is 5·40. The series expansion of this exponential and all other exponentials having a smaller numerical value for their exponent is converging after, at most, the sixth term. Thus equation (16) can be evaluated by integrating and evaluating the important terms in the expansion.

Thus, using the boundary condition $C = C_0$ at $x = 0$,

$$\frac{C_x - C_0}{C_0 - C_s} = -\int_0^X \left(1 - X + \frac{X^2}{2!} - \ldots\right) \times \times \left(1 - BX^3 + \frac{B^2X^6}{2!} - \ldots\right) dX \qquad (19)$$

Combining equations (13) and (19), we obtain the supercooling at any point x

$$S_x = -mC_s\frac{(1-k)}{k}\left\{(1-e^{-X}) - B\frac{X^4}{4} + B\frac{X^5}{5} - \frac{B}{2!} \times \times \frac{X^6}{6} + \left(\frac{B^2}{2!} + \frac{B}{3!}\right)\frac{X^7}{7} - \left(\frac{B^2}{2!} + \frac{B}{4!}\right)\frac{X^8}{8} + \left(\frac{B^2}{2!2!} + \frac{B}{5!}\right)\frac{X^9}{9} - \ldots\right\} - G_L x \qquad (20)$$

The maximum supercooling is obtained by evaluating X' from equation (18) and substituting it into equation (20).

To illustrate the dependence of the maximum supercooling on the parameters f and ω, a practical example has been evaluated. In Fig. 4 values of the maximum supercooling are plotted against f for three values of δ for the case of germanium doped with gallium (full lines). The value of C_s used (0·72 atomic per cent) was chosen such that $(dS/dx)_{x=0} = 0$ for $f = 0{\cdot}2 \times 10^{-3}$ cm sec^{-1}. Also, for convenience, the values of S_x have been calculated at a constant $G_L = 53{\cdot}2$°C/cm. Fig. 4 shows that, unlike $(dS/dx)_{x=0}$, the maximum supercooling depends on the stirring rate: i.e., at a constant value of $(dS/dx)_{x=0}$, the maximum supercooling

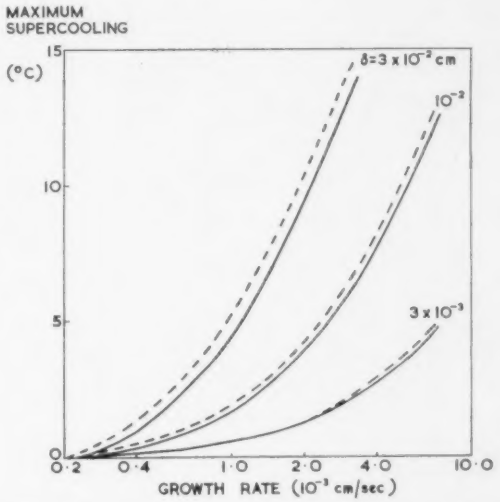


FIG. 4. Maximum supercooling vs. growth rate for three values of δ (0·72 atomic per cent gallium in germanium). (——— Accurate theory; – – – – approximate theory.)

will depend on the value of δ, being greater the greater the value of δ. It may be possible, under some conditions, for the maximum supercooling to reach a sufficiently large value that spontaneous nucleation occurs in the supercooled region thus leading to polycrystalline growth.

To show that the condition $X'/\Delta < 1{\cdot}84$ has not been violated in the above calculation, the values of X'/Δ are plotted against f in Fig. 5. As f is increased, X'/Δ tends to a limiting value which depends slightly on δ, and which is less than 1·84 for the three values of δ which were taken. X'/Δ becomes zero at $f = 0{\cdot}2 \times 10^{-3}$ cm sec^{-1} [where $(dS/dx)_{x=0} = 0$].

An approximate value for the maximum supercooling can be obtained by combining equations (3), (7) and (13) to give

$$S_x = -mC_s\frac{(1-k)}{k}(1-e^{-X}) - G_L x \qquad (21)$$

[This is in fact equation (20) with $B = 0$.]

On this approximate theory, the position of the maximum supercooling will be given by equation (14) where $(dC/dx)_{x=x'}$ is obtained from equation (2), provided $x' < \delta$ (line 2 of Fig. 2). Otherwise

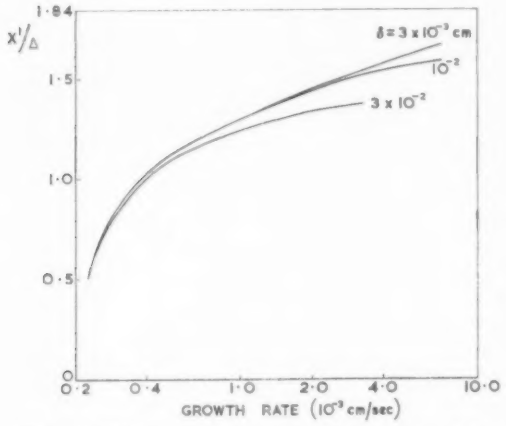


FIG. 5. X'/Δ vs. growth rate for three values of δ.

the position of the maximum supercooling will be at $x' = \delta$ (line 3 of Fig. 2). Since, in practice, the range of values of ω and f commonly used are such that the characteristic distance (D/f) of the exponential solute distribution in the diffusion-dominated zone is much greater than δ, the simple mathematical approximation to the solute distribution in the region $0 \leqslant x \leqslant \delta$ [given by equations (2) and (3)] is nearly linear. Hence, on this theory, for all but the smallest values of $(dS/dx)_{x=0}$, the maximum supercooling occurs at $x' = \delta$, and is

$$S_\delta = -mC_s \frac{(1-k)}{k}(1-e^{-\Delta}) - G_L\delta \qquad (22)$$

Values of S_δ are also plotted on Fig. 4 (dotted lines).

The agreement between the approximate and accurate values of maximum supercooling is good for large values of supercooling. The good agreement is a consequence of two facts:

(i) For reasonable values of G_L and for large supercoolings, $G_L x'$ and $G_L\delta$ are both much smaller than $-mC_s(1-k)(1-e^{-\Delta})/k$.

(ii) For large supercoolings

$$\left\{(1-e^{-X'}) - B\frac{X'^4}{4} + B\frac{X'^5}{5} \ldots .\right\} \simeq (1-e^{-\Delta})$$

This is equivalent to saying that $(C_{x'} - C_L)/(C_0 - C_L)$ is small.

DISCUSSION

The existence of constitutional supercooling during the growth of a metal crystal causes the solid at the solid–liquid interface to adopt a cellular form.[1] The boundaries of the cells are depressed with respect to the centres of the cells and are richer in solute than the centres of the cells (provided $k < 1$). The cells are perpetuated throughout the body of the crystal, and the crystal therefore contains "pipes" of solute-rich material which delineate the successive positions of the cells as the crystal was grown.

However the value of $(dS/dx)_{x=0}$ at which the transition from a planar to a cellular interface takes place is thought, by TILLER,[11] to be different in different solvents. BOLLING *et al.*[10] grew crystals of germanium doped with gallium in a horizontal boat and concluded that a large degree of constitutional supercooling was required before cellular formation occurred with this alloy. They also noted that lamellar twinning preceded cell formation.

TILLER[12] has postulated that arrays of dislocations form at the cell boundaries in order to relieve the stress which is produced, because the high solute concentration at the cell boundaries changes the lattice parameter there. It has been noted, in this laboratory, that a large increase in dislocation density accompanies cell formation in germanium crystals doped with gallium, which have been pulled from the melt.

The assumptions of one-dimensional thermal and solute flow made in the above analysis will be valid only for the case of a planar interface and stirring conditions such as to maintain a radial uniformity of solute. Once well-developed cells have formed, there will be a radial component to the solute flow which is manifested by the enhanced segregation to the cell boundaries. However, TILLER[11] has shown, at least for the case of a metal, that the formation of a cellular interface does not markedly alter the solute concentration in the liquid immediately ahead of the interface (i.e. it does not appreciably relieve the supercooling). Hence the above equations may not be seriously in error even when cells are present.

The assumption that the melt is infinite implies

that the concentration in the bulk melt (C_L) is constant. In any real crystal growth experiment, however, the concentration in the bulk melt is steadily increasing for solutes with $k<1$, in accordance with the "normal freeze" equation

$$C_L^g = C_L^0 (1-g)^{k_{eff}-1} \tag{23}$$

where g is the fraction of the crystal which has solidified. For the analysis given in this paper to be valid for this real case, the length of the "transient" region must be small compared with the length of the crystal. The length of the transient region depends on the height to which the solute builds up; with effective stirring the build-up is small. The amount of build-up is shown by the shaded area in Fig. 1. Since $\delta < D/f$, this can be approximated by a right-angled triangle of area $(C_0-C_L)\delta/2$. The amount of solute rejected from unit area of the crystal when the interface moves a distance x is approximately $C_L\ (1-k)x$, [provided $(1-k)\simeq(1-k_{eff})$], where C_L is expressed for unit volume. Ignoring diffusion from the zone ($x = 0$ to δ), the length of the transient region is therefore given approximately by

$$\frac{C_0-C_L}{C_L}\frac{1}{(1-k)}\delta/2 = \frac{k_{eff}-k}{k}\frac{1}{(1-k)}\delta/2$$

For $k = 0{\cdot}1$, $\delta = 10^{-2}$ cm and $\Delta = 0{\cdot}5$, the length of the transient region is approximately $0{\cdot}3\times10^{-2}$ cm. The fact that the BURTON *et al.* theory[8] accurately describes the solute distribution in pulled crystals provides evidence in support of the conclusion that the length of the transient region is short.

The value of m is the slope of the liquidus line at the concentration C_0. Comprehensive data for the liquidus curves of a large number of solute elements in germanium and silicon have recently been published by THURMOND and KOWALCHIK[13]. If the liquidus solution is "ideal", the equilibrium liquidus temperature (T^E) is related to the atom fraction of the *solvent* (y) present by[13]

$$\ln y = \frac{L}{R}\left(\frac{1}{T^0}-\frac{1}{T^E}\right) \tag{24}$$

where T^0 is the melting temperature of the pure solvent. From (24) the slope of the liquidus curve at a composition (y) is

$$\frac{dT^E}{dy} = \frac{L(T^0)^2R}{(L-T^0R\ln y)^2 y} \tag{25}$$

which reduces to $\dfrac{(T^0)^2R}{L}$ for very dilute solutions.

THURMOND and KOWALCHIK[13] show that this represents a very good approximation for the system gallium–germanium.

The analysis will not apply if the "facet effect"[14] is present (i.e. if a facet is formed on the solid–liquid interface on which the distribution coefficient has a different value).

From the point of view of achieving a high lattice perfection and solute uniformity, which are pre-requisites for device reproducibility, it can be seen that it is necessary to grow the crystals under conditions which avoid the occurrence of an appreciable degree of constitutional supercooling. Thus the growth conditions should be such that the magnitude of $(dS/dx)_{x=0}$ from equations (11) or (12), should be small or zero. For the incorporation of a given concentration of solute in the solid and for $(dS/dx)_{x=0} = 0$, the growth rate must be less than or equal to a critical value which can be obtained from Fig. 3.

An investigation of the effects of constitutional supercooling on the growth morphologies of germanium is currently being carried out in this laboratory.

Acknowledgements—The author is grateful to Drs. BARDSLEY, HULME and MULLIN for helpful discussions, and to the Controller, H.M. Stationery Office, for permission to publish.

REFERENCES

1. J. W. RUTTER and B. CHALMERS, *Canad. J. Phys.* **31,** 15 (1953).
2. W. A. TILLER, K. A. JACKSON, J. W. RUTTER and B. CHALMERS, *Acta Met.* **1,** 428 (1953).
3. W. A. TILLER and J. W. RUTTER, *Canad. J. Phys.* **34,** 96 (1956).
4. D. WALTON, W. A. TILLER, J. W. RUTTER and W. C. WINEGARD, *J. Metals, N.Y.* **7,** 1023 (1955).
5. T. S. PLASKETT and W. C. WINEGARD, *Canad. J. Phys.* **37,** 1555 (1959).
6. L. ESAKI, *Phys. Rev.* **109,** 603 (1958).
7. J. A. BURTON and W. P. SLICHTER, *Transistor Technology* Vol 1, Chaps. 5 and 6. Van Nostrand, New York (1958).

8. J. A. Burton, R. C. Prim and W. P. Slichter, *J. Chem. Phys.* **21,** 1987 (1953).
9. W. G. Cochran, *Proc. Camb. Phil. Soc.* **30,** 365 (1934).
10. G. F. Bolling, W. A. Tiller and J. W. Rutter, *Canad. J. Phys.* **34,** 234 (1956).
11. W. A. Tiller, *Growth and Perfection of Crystals* p.323 Wiley, New York (1958).
12. W. A. Tiller, *J. Appl. Phys.* **29,** 611 (1958).
13. C. D. Thurmond and M. Kowalchik, *Bell Syst. Tech. J.* **39,** 169 (1960).
14. K. F. Hulme and J. B. Mullin, *Phil. Mag.* **4,** 1286 (1959).

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 45–50. Printed in Great Britain

DIFFUSION OF CHARGED PARTICLES INTO A SEMICONDUCTOR UNDER CONSIDERATION OF THE BUILT-IN FIELD

KURT LEHOVEC and ALEXIS SLOBODSKOY

Sprague Electric Company, North Adams, Massachusetts

(*Received* 30 *January* 1961)

Abstract—The distribution of charged impurities in a semiconductor is calculated for the case of impurities diffusing from the surface into an infinite body, taking into account the electric field arising by the space distribution of the charged impurities. Fig. 2 shows the impurity distribution for different surface concentrations and Fig. 3 gives the true surface concentration as a function of the apparent surface concentration obtained by extrapolating from the tail of impurity distribution to the surface, using an ordinary complementary error function, i.e. ignoring the presence of a built-in field.

Résumé—La distribution d'impuretés chargées dans un semiconducteur est calculée pour le cas d'impuretés diffusant de la surface dans un corps infini, en tenant compte du champ électrique issu de la distribution de ces impuretés chargées dans l'espace.

La Fig. 2 démontre la distribution d'impuretés pour différentes concentrations de surface et la Fig. 3 donne la vraie concentration de surface en fonction de la concentration de surface apparente obtenue en extrapolant une fonction ordinaire d'erreur complémentaire de la base de la distribution d'impuretés à la surface, c'est à dire en ignorant la présence du champ developpé.

Zusammenfassung—Die Verteilung geladener Fremdatome in einem Halbleiter wurde berechnet, unter der Voraussetzung, dass die Fremdatome von der Oberfläche in einen unendlich grossen Körper eindiffundieren und unter Berücksichtigung des durch die Raumverteilung der geladenen Störstellen erzeugten elektrischen Feldes. Abb. 2 zeigt die Verteilung der Störstellen bei verschiedenen Oberflächenkonzentrationen und Abb. 3 gibt die tatsächliche Oberflächenkonzentration als Funktion der scheinbaren, die durch Extrapolation vom Ende der Störstellenverteilung zur Oberfläche unter Verwendung einer komplementären Fehlerfunktion erhalten wurde, d.h. man vernachlässigt das Vorhandensein eines eingebauten Feldes.

1. INTRODUCTION

DIFFUSION of impurities into semiconductors has attracted considerable attention in recent years. If the diffusion occurs from a gas phase, it is generally assumed that the concentration of impurities in the semiconductor near the surface is a function of the temperature during the diffusion and of the vapor pressure of the impurities in the gas phase, but does not change during the diffusion. Assuming further that the diffusion constant does not depend on the concentration of impurities, Fick's law of diffusion shows that the impurity distribution in a semi-infinite solid is a complementary error function. More recently it has been pointed out by ZAROMB[1], SMITS[2] and by KURTZ and YEE[3] that an electric field arises during the diffusion of electrically charged particles which enhances the motion of those particles leading to an impurity distribution differing from the complementary-error-function curve. Since the electric field arising from the charged particles is proportional to the gradient of the particle concentration,[2, 3] the field current may be considered simply as another diffusion current leading to a differential equation for the particle distribution, which is identical to Fick's law except that the diffusion "constant" is now a function of the particle concentration. In this

paper, we shall discuss the solution of this modified diffusion equation for the boundary conditions that the concentration of impurities at the surface is a constant and that the concentration of impurities at large distances from the surface is zero. The calculations of this paper apply to the case when all impurities are ionized. This is not a good approximation, even at zero ionization energy of the impurities, if the impurity concentration is sufficiently high that the holes become degenerate[4], and also the expression for the electric field differs from our equation (5). Our calculations remain valid for impurities of small ionization energies, if the impurity concentration is less than the partition sum of holes for acceptor-type impurities (or that of electrons for donor-type impurities).

2. MATHEMATICAL FORMULATION OF THE PROBLEM

We shall assume that we are dealing with an intrinsic semiconductor and that the diffusing impurities have a single negative charge. A significant electric field occurs during the diffusion of impurities into the semiconductor in the region where the impurity concentration exceeds the intrinsic carrier concentration. The reason for this field can be understood easily as follows.

Space-charge neutrality would require that

$$p = C+n \tag{1}$$

where C, p and n are the concentrations of impurities, holes and electrons. However, in the region where

$$C \gg n \tag{2}$$

one has from (1) $p \simeq C$, showing that the diffusion current of holes is a large multiple of the diffusion current of impurities, owing to the large ratio of hole mobility to impurity mobility

$$\mu_p/\mu_C \gg 1 \tag{3}$$

Since the total electric current during the diffusion must be approximately zero, the current due to holes must be equal but of opposite sign to the current due to impurities. Thus, the large diffusion current by holes must be just about compensated by a field current of holes flowing in the opposite direction. This requires an electric field which is opposing the diffusion of holes and thus accelerating the diffusion of the oppositely charged impurities. The magnitude of this electric field E has been given by KURTZ and YEE[3]. Their expression can be simplified considering equations (1) and (3) and using†

$$np = n_i^2 \tag{4}$$

to obtain

$$E = \frac{kT}{q} C^{-1} \frac{dC}{dx} \sqrt{\frac{(C/n_i)^2}{[(C/n_i)^2+4]}} \tag{5}$$

Note that the electric field is proportional to the impurity gradient.‡ Thus, the sum of field current and diffusion current of impurities becomes proportional to the gradient of the impurity concentration and may be considered as a diffusion current of the effective diffusion constant

$$D^* = D(1+f) \tag{6}$$

where

$$f = \frac{C/n_i}{[(C/n_i)^2+4]^{1/2}} \tag{7}$$

This effective diffusion constant was already given by SMITS[2] in his equation (7). Conservation of impurities requires that

$$\frac{\partial C}{\partial t} = \frac{\partial}{\partial x}\left(\frac{D^*\partial C}{\partial x}\right) \tag{8}$$

Introducing

$$u = \frac{x}{(2Dt)^{1/2}} \tag{9}$$

$$c = C/n_i \tag{10}$$

one obtains§ from (6), (7) and (8)

$$-\frac{dc}{du}2u = (1+f)\frac{d^2c}{du^2} + \frac{4}{(c^2+4)^{3/2}}\left(\frac{dc}{du}\right)^2 \tag{11}$$

† We are here concerned with such high temperatures that the intrinsic carrier concentration n_i is much larger than the bulk impurity concentration of the semiconductor.

‡ The presence of a space-dependent electric field requires space charges which appear to contradict equation (1). However, it can be shown, by calculating dE/dx from equation (5) and using the solution $C(x)$, that the magnitude of these space charges is negligibly small compared to C. Thus, equation (1) represents a valid approximation.

§ A possible dependence of the ordinary diffusion constant D on the impurity concentration will be neglected.

Dividing by

$$\frac{dc}{du}(1+f)$$

and integrating, one obtains

$$-\int \frac{2u\,du}{1+f} = \ln c' + \ln(1+f) \tag{12}$$

from which

$$c+(c^2+4)^{1/2}+k_1 = k_2 \int \exp\left(-\int \frac{2u\,du}{1+f}\right) du \tag{13}$$

where k_1 and k_2 are integration constants to be determined from the boundary conditions.

This form of the solution would be well suited for numerical computation, except that the function c appears in the term f on the right-hand side. However, the denominator containing this function varies only between 1 and 2 as c varies from zero to infinity. This suggests that a reasonable zero-order approximation to c by an explicit function of u substituted in f in equation (13) should provide a quite accurate first-order approximation.

3. CHOICE OF THE ZERO-ORDER APPROXIMATION

In the case of *indiffusion*, we have the boundary conditions

$$\left.\begin{aligned} c &= c_0 \text{ at } u = 0 \\ \text{and} \qquad c &= 0 \text{ at } u = \infty \end{aligned}\right\} \tag{14}$$

Accordingly, the zero-order approximation to (7) must be a function of u which varies from a value $\leqslant 1$ at $u = 0$ to zero at $u = \infty$. Furthermore, it is practical to select this function from a family for which the integral in equation (12) can be expressed by ordinary functions. These conditions are satisfied by the function

$$h = [1+\beta \exp\alpha(u^2-u^{*2})]^{-1} \tag{15}$$

for which

$$\exp\left(-\int \frac{2u\,du}{1+h}\right) =$$

$$= \text{const.} \exp(-u^2/2)[2\exp(\alpha u^{*2})+\beta\exp(\alpha u^2)]^{-1/2\alpha}$$

$$= \text{const. } S(u,u^*,\alpha,\beta) \tag{16}$$

Substituting (16) for the integral in (13) and considering the boundary conditions (14), one obtains

$$\frac{c+(c^2+4)^{1/2}-2}{c_0+(c_0^2+4)^{1/2}-2} = \frac{J(u,u^*,\alpha,\beta)}{J(0,u^*,\alpha,\beta)} \tag{17}$$

with

$$J(u,u^*,\alpha,\beta) = \int_u^\infty S(u,u^*,\alpha,\beta)\,du \tag{18}$$

Numerical values of the three parameters, α, β, u^* will be chosen in such a manner that at $u = u^*$

$$h = f \tag{19}$$

$$h' = f' \tag{20}$$

$$h'' = f'' = 0 \tag{21}$$

i.e. that our approximation h is self consistent with the function f calculated from the exact solution c in the vicinity of the inflection point. One recognizes easily that the vicinity of the inflection point is the region of largest importance for the equation (13). For, if $\beta \exp \alpha(u^2-u^{*2}) \gg 1$, the error in this expression effects little the accuracy of $1+h \simeq 1$, while if $\beta \exp\alpha\,(u^2-u^{*2}) \ll 1$ the error of this expression effects little the accuracy of $h \simeq 1$.† One obtains from $h'' = 0$ that

$$u^{*2} = \frac{1+\beta}{2\alpha(\beta-1)} \tag{22}$$

while from (19)

$$c_{u=u^*} = c^* = \frac{2}{\sqrt{[\beta(\beta+2)]}} \tag{23}$$

Equation (20) provides

$$\alpha = \frac{(1+\beta)^2}{3+\beta} \tag{24}$$

where c' occurring in h' has been eliminated by using

$$c'_{u=u^*} = -\frac{2u^*(c^{*2}+4)^{3/2}}{3c^*(1+f)(c^{*2}+4)^{1/2}+4} \tag{25}$$

† In the case that c_0 is *not* large compared to unity, it may be more advantageous to select the parameters from the conditions: $f = h$, $f' = h'$, $f'' = h''$ at the urface $u = u^* = 0$.

which comes from $h'' = 0$ after eliminating c'', using (11). Combining (22) and (24) provides

$$u^* = \left[\frac{\beta+3}{2(\beta^2+1)}\right]^{1/2} \tag{26}$$

The values of u^*, c^* and α are plotted as functions of the parameter β in Fig. 1.

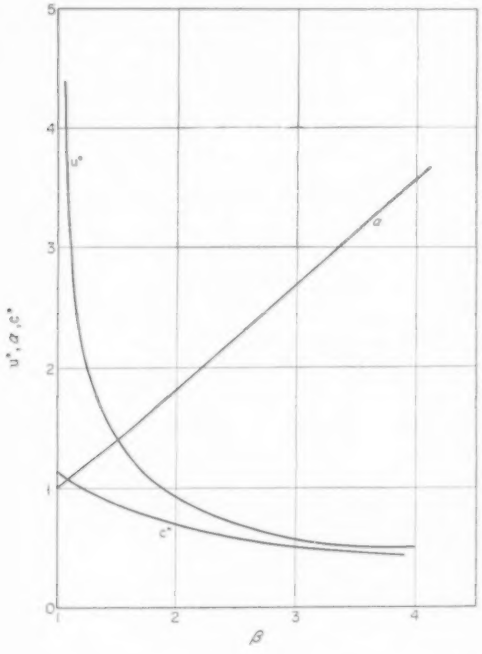


FIG. 1. The parameters u^*, α and c^* as functions of β according to equations (26), (24) and (23). These parameters inserted in equation (15) which is then substituted for f in (13), provide a self-consistent solution at the point $u = u^*$.

The integral of equations (18) has been evaluated numerically using an IBM computer for values of u^* ranging from 0·5 to 3·0 in increments of 0·5. Fig. 2 shows a family of distributions of impurities for various values of the parameter u^*. The dotted curve is an ordinary complementary-error-function curve fitted in the range $c \ll 1$ to the actual impurity distribution pertaining to $u^* = 2$ and extrapolated to the surface. This provides the "apparent surface concentration" c_a which can be compared to the actual surface concentration c_0. The relationship between c_a and c_0

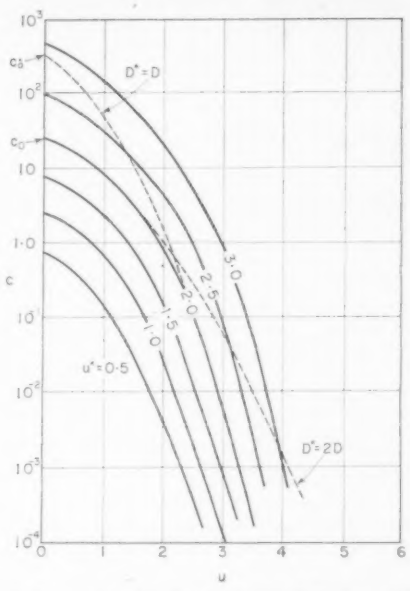


FIG. 2. Impurity distribution by indiffusion taking into account the built-in electric field. Ordinate is the impurity concentration divided by the intrinsic carrier concentration. Abscissa is the distance from the surface divided by the diffusion length. The tail of the curve pertaining to $u^* = 2$ has been fitted by a complementary-error-function curve which has been extended to the surface providing the apparent surface concentration c_a. Also shown is a complementary-error-function curve pertaining to the diffusion constant $2D$ and drawn through the true surface concentration c_0 of the curve pertaining to $u^* = 2$.

can be obtained mathematically as follows. For $\beta \exp(\alpha u^2) \gg 2 \exp(\alpha u^{*2})$ the integral (18) becomes

$$J(u,u^*,\alpha,\beta) \simeq \beta^{-1/2\alpha}\frac{\sqrt{\pi}}{2}\operatorname{erfc} u \tag{27}$$

The tail of the function (17) for $u \gg u^*$ and $c \ll 1$ is given by

$$\frac{c}{c_0+(c_0^2+4)^{1/2}-2} = \frac{\beta^{-1/2\alpha}\pi^{1/2}\operatorname{erfc} u}{2J(0,u^*,\alpha,\beta)} \tag{28}$$

Extrapolating from this tail to the surface one obtains the apparent surface concentration

$$c_a = \frac{\pi^{1/2}[c_0+(c_0^2+4)^{1/2}-2]\beta^{-1/2\alpha}}{2J(0,u^*,\alpha,\beta)} \tag{29}$$

This relation is plotted in Fig. 3, using different values of u^* as parameter and the corresponding values of α, β from equations (24) and (26).

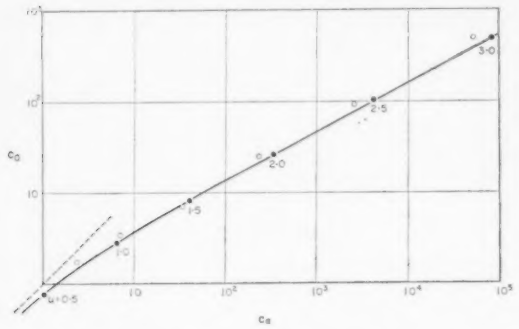


FIG. 3. True surface concentration c_0 vs. apparent surface concentration c_a. Both surface concentrations are expressed in reduced co-ordinates by dividing through the intrinsic-carrier concentration at the diffusion temperature. The curve plotted and the dots on this curve are obtained from the zero-order approximation, equation (15) substituted for f in equation (13) with the values of the parameters α, β functions of u^*, as shown in Fig. 1. The points ○ were obtained from a step function $h = 1$ for $u < u^*$, and $h = 0$ for $u > u^*$ as zero-order approximation to f in equation (13). The dotted line represents the relation $c_0 = c_a$ which is approached asymptotically for $c_0 \ll 1$.

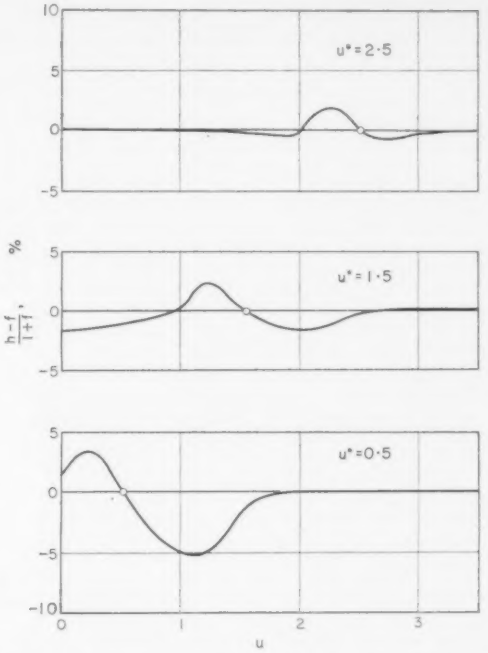


FIG. 4. The relative error $(h-f)/(1+f)$, where h is our zero-order approximation (15) and f is calculated from the corresponding first-order approximation using (7).

4. ACCURACY OF THE RESULTS

The accuracy of the mathematical solution (17) can be estimated by inspecting the self consistency of the zero-order approximation h, equation (15), with the value of f calculated from equation (7) using our first-order approximation plotted in Fig. 2. The expression $(h-f)/(1+f)$ which is the relative error of $1+f$, has been plotted in Fig. 4 for $u^* = 0{\cdot}5$, $1{\cdot}5$ and $2{\cdot}5$. Note that this error is zero for $u = u^*$, in accordance with the choice of the parameters u^*, α and β. The error vanishes for $u \gg u^*$, since h and f approach zero, while $1+f$ remains finite. For $u > u^*$, the error is negative. For u smaller than but closer to u^* the error is positive and becomes negative as u approaches zero. The maximum value of the error is 5 per cent for $u^* = 0{\cdot}5$ and less than 2 per cent for $u^* = 2{\cdot}5$. Because of the small magnitude of errors, and of the fact that the errors in various ranges of u have different signs, it is believed that the error in the curve c_a versus c_0 of Fig. 3 is at most a few per cent, and probably less than 1 per cent.

For comparison, a few values of c_a vs. c_0, arising from the step function $D^* = 2D$ for $u < u^*$ and $D^* = D$ for $u > u^*$, are also plotted in Fig. 3. For this case the term $\partial D^*/\partial x$ vanishes (except at $u = u^*$) and the differential equation (8) can be solved in closed form. In solving the equation we used that $c_{u^*-\epsilon} = c_{u^*+\epsilon}$ and $2Dc'_{u^*-\epsilon} = Dc'_{u^*+\epsilon}$ for $\epsilon \to 0$. The parameter u^* was chosen that $f = \frac{1}{2}$, i.e. $c^* = \sqrt{(4/3)}$ at $u = u^*$. This choice of D^* corresponds to the zero-order approximation $h = 1$ for $u < u^*$ and $h = 0$ for $u > u^*$. Note that even the crude zero approximation of f by a step function yields a relation c_a vs. c_0 of fair accuracy.

5. SUMMARY AND CONCLUSIONS

Solutions have been presented for the impurity distribution in a semiconductor arising from indiffusion, taking into account the electric field arising from the diffusing impurities. The solution

provides corrections to the apparent surface concentration which is obtained by extrapolating from the tail of the impurity distribution to the surface using a complementary error function, i.e. ignoring the electric field enhancing diffusion. An example for a method which determines the surface concentration by extrapolation from the tail of the diffusion curve is that based on the location of a *p–n* junction arising from indiffusing impurities into a sample having a bulk doping of the opposite conductivity. In such a case, and if the true surface concentration of impurities is large compared to the intrinsic carrier concentration at the diffusion temperature, the correction from the apparent surface concentration to the true surface concentration is very important, since otherwise errors by orders of magnitude would result. On the other hand, when determining surface concentrations from sheet-conductance measurements, the diffusion length enters merely as a proportionality factor, and the enhancement of the effective diffusion constant by the electric field can introduce an error by not more than a factor $\sqrt{2}$.

Acknowledgement—We gratefully acknowledge the assistance of Mr. KURT SCHOENI in carrying out the numerical integrations on an IBM computer.

REFERENCES

1. S. ZAROMB, *IBM J. Res. Dev.* **1,** 57 (1957).
2. F. M. SMITS, *Proc. I.R.E.* **46,** 1049 (1958).
3. A. D. KURTZ and R. YEE, *J. Appl. Phys.* **31,** 303 (1960).
4. J. W. ALLEN, *J. Phys. Chem. Solids* **15,** 134 (1960).

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 51–58. Printed in Great Britain

BARIUMTITANAT ALS SPERRSCHICHTHALBLEITER*

W. HEYWANG

Forschungslaboratorium der Siemens und Halske AG., München 8

(*Received* 14 *February* 1961)

Zusammenfassung—In geeignet dotierter ferroelektrischer Titanatkeramik—insbesondere in $BaTiO_3$—wird in einem beschränkten Temperaturbereich unmittelbar oberhalb des Curiepunktes θ ein anomaler Abfall der Leitfähigkeit mit steigender Temperatur um ca. 4 Zehnerpotenzen beobachtet. In diesem Temperaturbereich sinkt der Widerstand mit zunehmender Feldstärke, und zwar um so steiler, je höher er mit der steigenden Temperatur geworden ist. Dieses Verhalten weist darauf hin, dass der hohe Widerstand auf Sperrschichten konzentriert ist, die sich auf Grund von Oberflächentermen an den Korngrenzen im Innern der Keramik ausbilden. Mit dieser Deutung steht auch die Höhe der dem Widerstand parallel liegenden Kapazität der Proben in Einklang.

Die oberhalb θ stark temperaturabhängige Dielektrizitätskonstante bestimmt wesentlich den Verlauf des elektrostatischen Potentials innerhalb der Sperrschichten und damit die Höhe des Potentialwalles an den Korngrenzen. Da diese exponentiell in den Ausdruck für die Leitfähigkeit der Sperrschichten eingeht, ist auch eine starke Temperaturabhängigkeit des Widerstandes zu erwarten, und zwar im Sinne eines Widerstandsanstiegs mit steigender Temperatur! Der Widerstandsanstieg im Bereich des Abfalls der DK oberhalb θ wird dadurch begrenzt, dass der Potentialwall nicht höher werden kann als die Aktivierungsenergie der Oberflächenterme. Ist diese erreicht, so sinkt bei weiterer Temperaturerhöhung der Widerstand wieder exponentiell ab, wie bei jedem normalen Halbleiter.

Die theoretische Behandlung gemäss dem vorgeschlagenen Modell führt zu einer Deutung des gesamten physikalischen Erscheinungsbildes und zu befriedigender quantitativer Übereinstimmung mit dem Experiment.

Abstract—In suitably doped ferroelectric Titanium ceramic—especially $BaTiO_3$—an abnormal decrease of conductivity with rising temperature (of the order of 10^4) is observed in a limited temperature range immediately above the curie point θ. In this temperature range, resistance decreases with increasing field strength, and the decrease becomes steeper as the resistance grows with rising temperature. This behaviour indicates that the high resistance is concentrated in the blocking-layers which develop at the grain boundaries within the ceramic as a result of the surface terms. This explanation is in agreement with the high value of the specimen capacitance which is in parallel with the resistance.

The dielectric constant, which above θ is strongly temperature dependent, essentially determines the distribution of electrostatic potential within the blocking layer, and thus the value of the potential barrier at the grain boundaries. As this constant enters exponentially into the conductivity expression of the blocking layers, a strong temperature dependence of the resistance is to be expected, so that resistance increases with rising temperature. The resistance increase in the region of the decrease of the d.c. above θ is limited by the fact that the potential barrier cannot become higher than the activation energy of the surface terms. If this value is reached the resistance begins to decrease exponentially with rising temperature, as with all ordinary semiconductors.

Theoretical treatment of the proposed model leads to an interpretation of the physical processes involved and achieves good agreement with the experiment.

Résumé—Dans le céramique ferroélectrique au titanium—spécialement l' O_3BaTi—convenablement dopé, on a observé une chute anormale de conductibilité en fonction de la hausse de température (environ 10^4) dans la gamme de température limitée juste au-dessus du point curie θ. Dans cette gamme de température la résistance décroît en fonction de l'augmentation de l'intensité du champ, et la chute est plus prononcée dans la région où la résistance augmente en fonction da la

* Erweiterte Fassung eines Vortrages auf der Tagung des Verbandes Deutscher Physikalischer Gesellschaften 18.10.60, Wiesbaden.

température. Ce comportement démontre que la haute resistance est concentrée dans les couches obstruantes qui se développent aux bornes du grain dans le céramique résultant des termes de la surface. Cette preuve est en accord avec la capacité élevée du specimen qui est en parallèle avec la résistance.

La constante diélectrique qui au-dessus de θ est en dépendence accrue de la température détermine principalement la distribution du potentiel électrostatique dans les couches obstruantes et ainsi la valeur de la barrière potentielle aux bornes du grain. Comme cette constante est un terme exponentiel de l'expression de conductibilité des couches obstruantes, on s'attend à ce que la résistance dépende fortement de la température. L'augmentation de la résistance dans la région de la chute du courant continu au-dessus de θ est limitée par le fait que la barrière potentielle ne peut pas être supérieure à l'énergie d'activité des termes de la surface. Si cette valeur est atteinte, la résistance commence à décroître exponentiellement en fonction de la hausse de température tout comme dans les semi-conducteurs ordinaires.

Le traitement théorique du modèle propose mène a une interprétation des procédés physiques opérants et est en accord avec les données expérimentales.

1. EXPERIMENTELLES TATSACHENMATERIAL

BARIUMTITANAT ist bekannt geworden wegen seiner ferroelektrischen Eigenschaften. Diese seien, so weit sie für das folgende von Bedeutung sind, in Abb. 1 kurz zusammengefasst.[1] Bei hohen Temperaturen (oberhalb 120°C) hat $BaTiO_3$ eine ungewöhnlich hohe Dielektrizitätskonstante, die mit sinkender Temperatur gemäss einem Curie–Weissschen Gesetz ansteigt

$$\epsilon = \frac{C}{T-\theta} \qquad \text{(I, 1)}$$

mit der Curiekonstanten $C = 1{,}2 \times 10^5$ und der Curietemperatur $\theta = 110°C$.

Kurz vor Erreichen des Wertes $\epsilon = \infty$ an der Curietemperatur θ [gemäss Gleichung (I,1)] bildet sich in den $BaTiO_3$–Kristalliten eine permanente Polarisation P in [100]–Richtung aus. Dadurch sinkt die Dielektrizitätskonstante für kleine Felder,

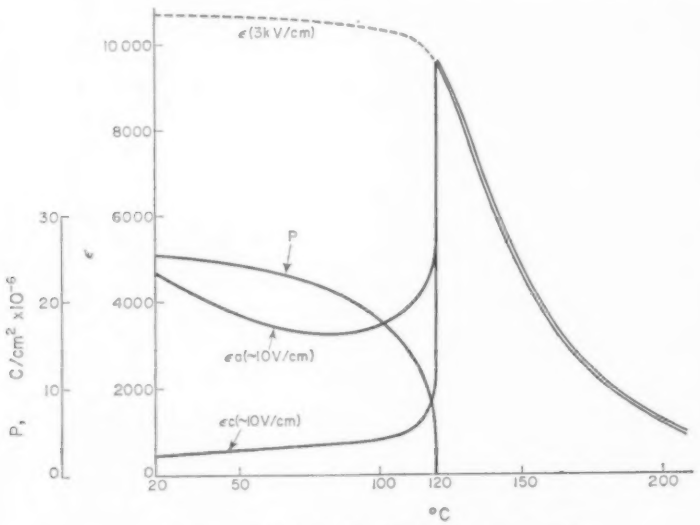


ABB. 1. Dielektrisches Verhalten von $BaTiO_3$.

abhängig von der Kristallrichtung, mehr oder weniger, ab (ϵ_c parallel, ϵ_a quer zur permanenten Polarisation). Durch das Zusammenwirken dieser anisotropen Dielektrizitätskonstanten mit Umklappvorgängen der permanenten Polarisation ergibt sich in einem polykristallinen Körper ein ziemlich kompliziertes dielektrisches Verhalten. Wichtig ist für das weitere jedoch nur, dass sich für grosse Felder (z.B. bei 3 kV/cm) eine sehr

hohe effektive Dielektrititäskonstante ergibt, die unterhalb θ von der Temperatur nur wenig abhängt (Abb. 1, gestrichelte Kurve).

Die Frage, ob sich der starke Temperaturgang der Dielektrizitätzkonstante auch auf eine—gleich aus welchen Gründen—vorhandene Leitfähigkeit auswirkt, haben sich erstmals BUSCH *et al.*[2] gestellt. Sie fanden bei schwach leitenden Proben eine kleine Anomalie in der Umgebung der Curiespitze (Abb. 2), die sie nach dem Mott-Gerneyschen Störstellenmodell als eine Rückwirkung der Dielektrizitätskonstante auf die

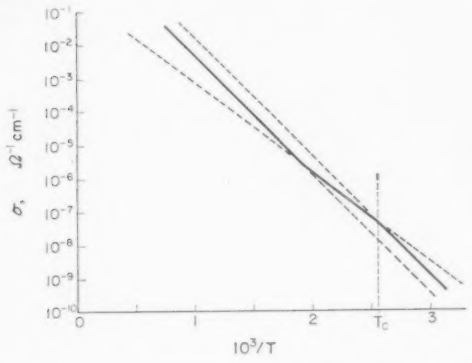


ABB. 2. Temperaturgang der Leitfähigkeit von $BaTiO_3$.

Aktivierungsenergie der Störstellen deuteten. Mit grösseren Effekten war auch kaum zu rechnen, nachdem MEYER[3] bereits bei früheren ausgedehnten Untersuchungen über die Leitfähigkeit anreduzierter Titanate keine Anomalien gefunden hatte, die einen Hinweis auf den—damals noch unbekannten—ferroelektrischen Charakter von $BaTiO_3$ gegeben hätten.

Überraschenderweise fanden nun HAAYMAN *et al.*[4] bei verschiedenen Titanaten einen hohen anomalen Widerstandsanstieg in einem begrenzten Temperaturbereich. Die Untersuchungen über die im folgenden berichtet werden soll[5], dienen der Aufgabe, die Natur dieses "Kaltleiter"–Effektes zu klären.

Als erstes zeigte sich, dass der starke anomale Widerstandsanstieg nur bei Titanatkeramik auftrat, die gemäss dem Prinzip der gelenkten Valenz dotiert war (z.B. bei teilweiser Substitution der Ba^{2+}-Ionen durch La^{3+}-Ionen oder der Ti^{4+}-Ionen durch Sb^{5+}). Er wurde aber nicht gefunden bei Keramik, deren Leitfähigkeit durch Anreduzieren erzeugt war, im Einklang mit den Untersuchungen Wilfried Meyers. Auch bei einer Dotierung nach dem Prinzip der gelenkten Valenz verschwand der Effekt mehr und mehr, sowie zusätzlich anreduziert wurde.

Untersucht wurden neben reinem $BaTiO_3$ auch Proben, deren Curiepunkt durch Substitution des Ba durch Sr bzw. des Ti durch Zr oder Sn verschoben war (Abb. 3). Dabei zeigte sich, dass der Widerstandsanstieg systematisch am Curiepunkt

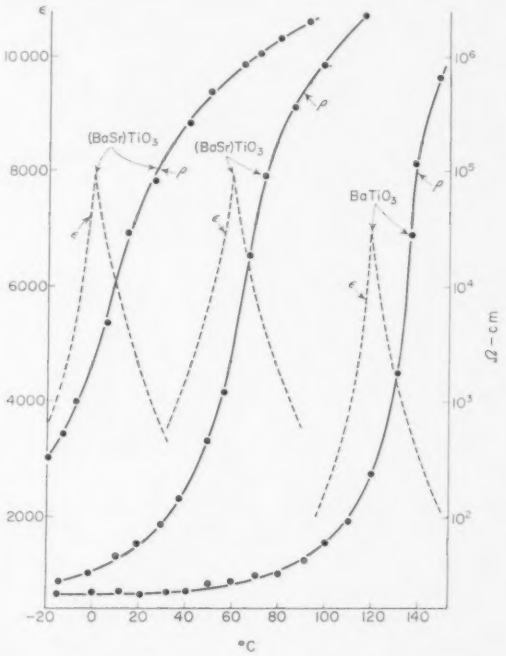


ABB. 3. Kaltleitereffekt in Titanatkeramik.

beginnt und sich etwa über einen Bereich von 100°C fortsetzt. Ein Deutungsversuch durch die am Curiepunkt eintretende schwache Gitterverzerrung schied wegen der grossen Ausdehnung des Anstiegsbereiches aus. Auffällig war jedoch, dass die Temperaturkurve des spez. Widerstands—dargestellt in logarithmischem Massstab—enge Verwandtschaft mit dem Temperaturgang der Dielektrizitätskonstanten für hohe Felder aufwies. Es schien daher folgender Ansatz zur

Deutung des anomalen Widerstandsanstiegs naheliegend

$$\sigma \simeq e^{-E(\epsilon)/kT} \qquad \text{(I, 2)}$$

wobei die Aktivierungsenergie E—im Gegensatz zu Heissleitern—infolge ihrer Abhängigkeit von ϵ stark temperaturabhängig ist.

Ein weiterer Einblick in die Natur des Effektes ergab sich durch Untersuchung der Abhängigkeit der Poissonschen Differentialgleichung

$$\Delta\phi = -\rho/\epsilon\epsilon_0 \qquad \text{(I, 3)}$$

verschwindend klein. Sie wächst aber stark an im Bereich des Curie–Weissschen Gesetzes, so dass nunmehr die Ladungsträger beim Stromdurchgang einen Potentialwall überschreiten müssen. Dadurch sinkt die Leitfähigkeit entsprechend

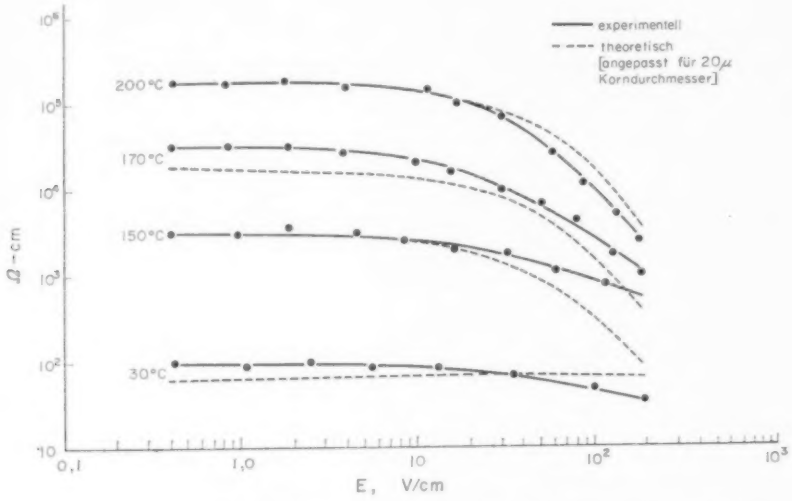


Abb. 4. Spannungsabhängigkeit des Kaltleiterwiderstandes Experiment und Theorie.

des Widerstandes von der angelegten Spannung (Abb. 4). Mit zunehmender Spannung sinkt nämlich der Widerstand, und zwar umso steiler, je höher er mit zunehmender Temperatur geworden ist. Eine derartige Spannungsabhängigkeit des Widerstandes tritt bekanntlich bei Sperrschichten auf, so dass man zu dem in Abb. 5(a) dargestellten Modell gelangt.* Infolge von Oberflächentermen bildet sich an der Grenze zwischen zwei Keramikkristalliten eine Raumladungszone aus, die eine Aufbäumung der Bänder bewirkt. Im Bereich hoher Dielektrizitätskonstanten—unterhalb der Curietemperatur—ist diese Aufbäumung gemäss Gleichung (I,2) exponentiell ab; denn man erhält mit $E \simeq 1/\epsilon$ und Gleichung (1):

$$\sigma \simeq e^{-\alpha/\epsilon T} = e^{-\alpha(T-\theta)/CT} \simeq e^{+\alpha\theta/CT} \qquad \text{(I, 4)}$$

(α = konst.)

Dies ergäbe einen mit der Temperatur stetig ansteigenden Widerstand; doch kann der Potentialberg mit abnehmender Dielektrizitätskonstante nicht beliebig anwachsen. Die maximale Höhe ist vielmehr erreicht, sobald die Oberflächenterme in den Bereich der Fermikante angehoben sind; denn bei weiterem Anheben werden die Oberflächenterme entleert und die Breite der Raumladungszone nimmt ab. Sowie die maximale Höhe des Potentialberges erreicht ist, sinkt der Widerstand bei weiterem Temperaturanstieg entsprechend dem üblichen Verhalten der Halbleiter ab.

Im folgenden sollen die hier skizzierten Ge-

* Das Bändermodell kann in diesem Fall auch für den oxydischen Halbleiter $BaTiO_3$ angewandt werden, da nicht die Feinheiten des Leitungsmechanismus behandelt werden, sondern vor allem die Verhältnisse in Raumladungszonen, die vom speziellen Modell unabhängig sind.

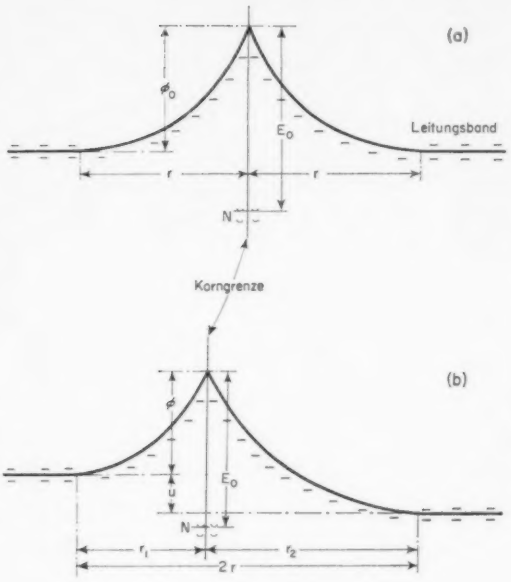


Abb. 5. Kaltleiter Randschichtmodell.

dankengänge quantitativ behandelt und mit dem Experiment verglichen werden.

2. QUANTITATIVE THEORETISCHE BEHANDLUNG

A. *Temperaturgang der Leitfähigkeit*

Im Volumen der Bariumtitanatkörner seien n_D Donatoren ins Gitter eingebaut, die infolge ihrer geringen Aktivierungsenergie quantitativ dissoziiert sind. Aus einer Randschicht der Breite r an den Kornoberflächen sind die Elektronen jedoch in die Oberflächenterme übergetreten, was eine Potentialaufbäumung des Leitungsbandes zur Folge hat (vgl. Abb. 5). Hierbei gelten folgende Zusammenhänge

$$rn_D = N_b \qquad \text{(II, 1)}$$

(N_b Dichte der besetzten Oberflächen-Terme, N Gesamtdichte der Oberflächenterme.)

$$\phi_0 = \frac{e^2 n_D}{2\epsilon\epsilon_0} r^2 \qquad \text{(II, 2)}$$

Bezeichnet man den Abstand der Oberflächenterme vom Leitungsband mit E_0, so gilt für ihren Besetzungszustand

$$N_b = \frac{N}{\exp[(\phi_0 - E_0 - \zeta)/kT] + 1} \qquad \text{(II, 3)}$$

Dabei ist die Lage der Fermikante ζ im Nicht-Entartungsfall bestimmt durch

$$n_D = N_c\, e^{\zeta/kT} \qquad \text{(II, 4)}$$

(N_c Bandgewicht des Leitungsbandes.)
Einsetzen dieses Ausdrucks in (II, 3) ergibt

$$N_b = N \Big/ \left(\frac{N_c}{n_D} e^{(\phi_0 - E_0)/kT} + 1 \right) \qquad \text{(II, 5)}$$

Durch Zusammenfassung der Gleichungen (II,1), (II,2) und (II,5) erhält man schliesslich die Bestimmungsgleichung für ϕ_0

$$\phi_0 = \frac{e^2}{2\epsilon\epsilon_0 n_D} \frac{N^2}{[(N_c/n_D) e^{(\phi_0 - E_0)/kT} + 1]^2} \qquad \text{(II, 6)}$$

Unter der bereits oben gemachten Voraussetzung einer vollen Dissoziation der Donatoren gilt für die Leitfähigkeit im Korninnern

$$\sigma_v = e\mu n_D \qquad \text{(II, 7)}$$

(μ Beweglichkeit).

Wegen der grossen Zahl der Korngrenzen und des—in Anbetracht der hohen Dielektrizitätskonstanten—flachen Verlaufs der Potentialberge kann angenommen werden, dass nur die Ladungsträger zur Leitfähigkeit beitragen, die den Potentialberg rein thermisch überwinden. Für die mittlere Leitfähigkeit der gesamten Keramik erhält man hiermit

$$\sigma = \sigma_v e^{-\phi_0/kT} \qquad \text{(II, 8)}$$

Wegen der geringen Temperaturabhängigkeit von σ_v ist für den Temperaturgang der Leitfähigkeit der Ausdruck ϕ_0/kT massgebend

$$\ln \sigma \simeq -\frac{\phi_0}{kT} = -\frac{e^2 N^2}{2C\epsilon_0 n_D k} \cdot \left(1 - \frac{\Theta}{T}\right) \Big/ \left(\frac{N_c}{n_D} \times \right.$$

$$\left. \times\, e^{\phi_0/kT - E_0/kT} + 1 \right)^2 \qquad \text{(II, 9)}$$

Damit ist ein Ausdruck für den Temperaturgang

der Leitfähigkeit im Bereich oberhalb der Curietemperatur gewonnen. Vor der weiteren Auswertung ist es jedoch zweckmässig, die Frage der Spannungsabhängigkeit zu verfolgen.

B. *Spannungsabhängigkeit des Widerstandes*

Zur Berechnung der Spannungsabhängigkeit sei vorausgesetzt, dass durch die Spannungsbelastung die Oberflächenterme nicht nennenswert umbesetzt werden. Die Randschicht zwischen zwei Körnern muss praktisch den gesamten auf die beiden Körner treffenden Spannungsabfall U aufnehmen. Dies führt zu einer Verschiebung der Raumladungszone. Hierfür gelten die Beziehungen [vgl. Abb. 5(b)]

$$r_1+r_2 = 2r \qquad \text{(II, 10)}$$

und

$$\frac{e^2 n_D}{2\epsilon\epsilon_0}(r_2^2-r_1^2) = U \qquad \text{(II, 11)}$$

oder

$$r_2-r_1 = \frac{U\epsilon\epsilon_0}{e^2 n_D r} \qquad \text{(II, 12)}$$

Für den Übertritt der Elektronen von einem Korn ins andere ist jetzt nicht mehr der Potentialberg ϕ_0, sondern der niedrigere Potentialberg ϕ massgebend. Seine Höhe ergibt sich aus

$$\phi = \frac{e^2 n_D}{2\epsilon\epsilon_0} r_1^2 = \frac{e^2 n_D}{2\epsilon\epsilon_0}\left(r-\frac{U\epsilon\epsilon_0}{2e^2 n_D r}\right)^2 \qquad \text{(II, 13)}$$

$$\phi = \phi_0\left(1-\frac{U}{4\phi_0}\right)^2$$

wobei vorausgesetzt werden muss, dass $U < 4\phi_0$ ist. Für grössere Spannungen verschwindet der Potentialberg völlig und der Kornwiderstand wird massgebend.

Für die Feldabhängigkeit der Leitfähigkeit gilt daher

$$\ln\sigma = \ln\sigma_v - \frac{\phi_0}{kT}\left(1-\frac{U}{4\phi_0}\right)^2 \quad U<4\phi_0 \qquad \text{(II, 14)}$$

$$\sigma = \sigma_v \qquad U>4\phi_0$$

3. AUSWERTUNG

Die transzendente Gleichung (II,9) muss numerisch ausgewertet werden. Dazu wurden ausser den Grundgrössen des Curie–Weissschen Gesetzes [Gleichung (I,1)] für $BaTiO_3$ folgende experimentellen Werte verwendet

$$\rho_v = \frac{1}{\sigma_v} = 30\ \Omega\text{-cm} \qquad \text{(III, 1)}$$

entsprechend dem spezifischen Widerstand bei Zimmertemperatur, und aus dem Widerstandsanstieg um 4 Zehnerpotenzen

$$\rho_{max}/\rho_v = 10^4$$
$$T_{max} = 240^\circ\text{C} \qquad \text{(III, 2)}$$

Des weiteren wurde angesetzt:
für das Bandgewicht

$$N_c = 2\times 10^{19}\ \text{cm}^{-3} \qquad \text{(III, 3)}$$

wie es sich für $m_{eff} = m_0$ (freie Elektronenmasse) ergibt, und für die unbekannte Beweglichkeit ein Bereich zwischen

$$\mu = 1 \text{ und } 300\ \text{cm}^2/\text{V-sec} \qquad \text{(III, 4)}$$

Aus μ und ρ errechnet sich eine Dichte der wirksamen Donatoren zwischen 3×10^{14} und 2×10^{17} cm^{-3}. Man erhält damit durch Anpassung an die Werte im Widerstandsmaximum [Gleichung (III, 2)] folgende Wertebereiche für Anzahl und Aktivierungsenergie der Oberflächenterme

$$N = 4\times 10^{11} \text{ bis } 10^{13}\ \text{cm}^{-2} \qquad \text{(III, 5)}$$

und

$$E_0 = 0{,}7 \text{ bis } 1{,}0\ \text{eV} \qquad \text{(III, 6)}$$

Der Wertebereich für N ist ebenso wie der von n_D sehr gross. Hingegen ist E_0 weitgehend durch den maximalen Widerstandsanstieg festgelegt. Das Ergebnis hinsichtlich des Temperaturganges ist in Abb. 6 wiedergegeben. Es zeigt sich eine im Hinblick auf die in der Theorie gemachten Vereinfachungen befriedigende Übereinstimmung mit dem experimentellen Kurvenverlauf und man erkennt, dass die Wahl von E_0 keinen grossen Einfluss hat.

Aus diesem Grunde beschränken sich die weiteren Rechnungen auf einen Mittelwert von $E_0 = 0{,}9$ eV. Unter Verwendung dieses Wertes ergibt sich der bereits in Abb. 4 eingetragene

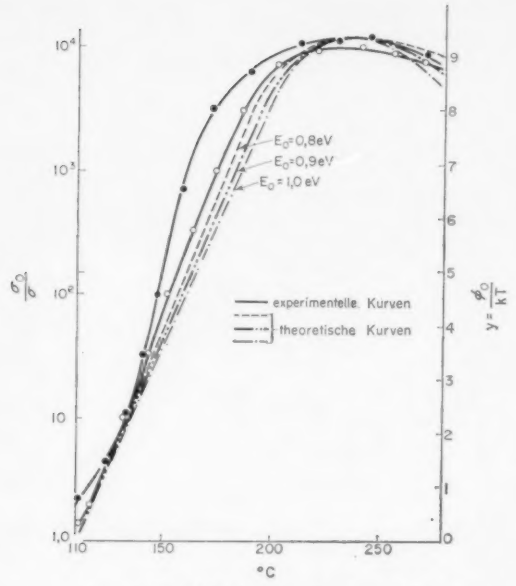


ABB. 6. Kaltleiter Widerstandsanstieg.

theoretische Verlauf der Spannungsabhängigkeit des Widerstandes. Zum Vergleich mit dem Experiment ist zu bemerken, dass die Spannungsabhängigkeit—ausser durch das Verhalten der einzelnen Sperrschicht, wie es aus der Rechnung folgt—noch durch die Anzahl der je cm hintereinander geschalteten Sperrschichten bestimmt wird. Eine optimale Anpassung der theoretischen an die experimentellen Kurven ergab sich für fünfzig Sperrschichten je cm in guter Übereinstimmung mit der direkt ermittelten Korngrösse von ca. 40 μ.

Des weiteren ergibt sich aus dem Modell die Dicke der Sperrschicht an den einzelnen Korngrenzen. Man erhält im oben betrachteten Fall eine Dicke von $2 \times 3{,}3\ \mu = 6{,}6\ \mu$, Die Feldstärke im Raumladungsgebiet erreicht damit die Grössenordnung von 10^3 V/cm, d. h. es musste die effektive Dielektrizitätskonstante für grosse Felder den Überlegungen zu Grunde gelegt werden. Dieser Wert ist um den Faktor 6 kleiner als der Korndurchmesser, d. h. die Sperrschichten greifen im Einklang mit den der Theorie zugrunde gelegten Voraussetzungen nicht ineinander über. Durch die Dicke der Sperrschichten ist aber auch die Kapazität C_s bestimmt, die parallel zum Korngrenzenwiderstand liegt, entsprechend dem Ersatzschaltbild der Abb. 7.

Dieses gilt zunächst nur für ein Einzelkorn, kann aber—wie die Frequenzabhängigkeit des Scheinwiderstandes zeigt—auch auf polykristallines Material im wesentlichen angewendet werden. C_s kann im Bereich des Widerstandsanstiegs, wo R_s gross ist gegen R_v, als kapazitiver Anteil des Scheinwiderstandes gemessen werden. Bei Proben

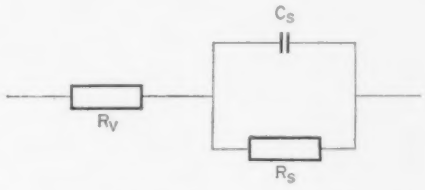


ABB. 7. Ersatzschaltbild: R_v Kornwiderstand, R_s Sperrschichtwiderstand, C_s Sperrschichtkapazität.

mit einer Korngrösse von 40 μ und einer Dotierung, wie sie der bisherigen Rechnung zugrunde gelegt wurden, liegt die aus C errechnete mittlere Dielektrizitätskonstante um den Faktor 10 über der Volumen-Dielektrizitätskonstanten des entsprechenden undotierten Körpers. Dies steht wiederum in guter Übereinstimmung mit dem oben ermittelten Verhältnis 6 zwischen Korndurchmesser und Sperrschichtdicke.

Nach diesen Ergebnissen darf das zur Deutung des anomalen Widerstandsanstiegs bei $BaTiO_3$ vorgeschlagene Sperrschichtmodell als gesichert angesehen werden.

Schliesslich sei noch folgendes bemerkt: Die bereits oben erwähnte Tatsache, dass anreduzierte Keramik höchstens einen verschwindend kleinen Widerstandsanstieg zeigt, ergibt sich zwanglos aus diesem Modell. Aus dem Temperaturgang des Widerstands unterhalb θ folgt nämlich bei schwach anreduzierten Proben—ebenso wie bei den Proben von Busch und Mitarbeitern (vgl. Abb. 2)—dass die Donatoren hohe Aktivierungsenergie besitzen. Eine Umbesetzung der Oberflächenterme, wie sie für den Kaltleitereffekt Voraussetzung wäre, kann also nicht auftreten.

Anerkennungen—Besonders danken möchte ich meinen Mitarbeitern Herrn FENNER für die Herstellung der Proben, Herrn Dr. SCHÖFER für phys. Messungen und Herrn SPICHTINGER für numerische Berechnungen.

ANHANG ZUR NUMERISCHEN BEHANDLUNG

Für die numerische Berechnung führt man zweckmässig in Gleichung (II,9) dimensionslose reduzierte Grössen ein

$$\frac{\theta}{T} = x;\ \frac{\phi_0}{kT} = y$$

$$\frac{e^2N^2}{2C\epsilon_0 n_D k} = A \qquad \text{(A, 1)}$$

$$\frac{N_c}{n_D} = e^b;\ \frac{E_0}{k\theta} = c$$

Für die Leitfähigkeit σ ist die Grösse y massgebend, für die nach Gleichung (II,9) unter Verwendung von Gleichung (A,1) der Zusammenhang gilt

$$y = \frac{A(1-x)}{[\exp(b+y-cx)+1]^2} \qquad \text{(A, 2)}$$

Aus der Bedingung $dy/dx = 0$ im Widerstandsmaximum erhält man folgenden Zussammenhang (Maximumswerte durch Index m gekennzeichnet)

$$\left.\begin{aligned} e^{b+y_m-cx_m}\,[1-2c(1-x_m)]+1 &= 0 \\ y_m &= A(1-x_m)\left[\frac{2c(1-x_m)-1}{2c(1-x_m)}\right]^2 \end{aligned}\right\} \qquad \text{(A, 3)}$$

Diese Gleichungen liefern einen Zusammenhang von der Art

$$\left.\begin{aligned} b &= b(c) \text{ und } A = A(c) \text{ bzw} \\ n_D &= n_D(E_0) \text{ und } N = N(E_0) \end{aligned}\right\} \qquad \text{(A, 4)}$$

wobei der Wertebereich für n_D bekannt ist. Hieraus ergeben sich für N und E_0 die in Gleichung (III,4) angegebenen Grenzen. Nach Wahl von E_0 in diesen Grenzen kann aus Gleichungen (A, 4) und (A, 2) der Temperaturgang des Widerstandes gewonnen werden.

LITERATUR

1. vgl. z.B. McQuarrie, *Bull. Amer. Ceram. Soc.* **34,** 256 (1955).
2. G. Busch, H. Flury und W. Merz, *Helv. phys. acta* **21,** 212 (1948).
3. W. Meyer, *Deut. Pat.* 660971 (1938).
4. P. W. Haayman, R. W. Dam und H. A. Klasens, *Deut. Pat.* 929350 (1955).
5. vgl. auch W. Heywang, *Solid State Physics in Electronics and Telecommunication* Vol. 4, p. 877. Academic Press, London (1960).

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 59–69. Printed in Great Britain

THE FORWARD TRANSIENT BEHAVIOR OF SEMICONDUCTOR JUNCTION DIODES

WEN-HSIUNG KO

Case Institute of Technology, Cleveland, Ohio

(*Received* 21 *November* 1960; *in revised form* 20 *March* 1961)

Abstract—The transient behavior of junction diodes in response to a step current in the forward direction is analyzed and the theoretical results verified experimentally. The forward transient of the diode is resolved into junction voltage and bulk resistance components. The former is due to the gradual building up of the excess carrier density near the junction and the latter due to the conductivity modulation.

A simplified planar junction model and diffusion equation are used in determining the excess carrier density as a function of time and position. From the excess carrier density function the two transient components may be determined analytically.

The theory yields satisfactory explanation of all the observed phenomena: the diode acts inductively for large forward current, capacitively for small current and oscillatory for certain intermediate current. Furthermore, these equations are confirmed with an accuracy of 5 per cent or better by experimental data from germanium and silicon alloyed diodes.

Several applications are suggested, for example the design methods to minimize the forward transient and the measurements of lifetime and bulk resistance from the transient observation.

Résumé—La manière de comportement transitoire des diodes à jonction en réponse à un courant à échelon dans la direction avant est analysée et les résultats théoriques vérifiés expérimentalement. La tension transitoire "avant" de la diode est résolue en une composante de voltage de jonction et une autre de résistance "en vrac". La première est due à la composition graduelle de l'excédent de la densité de porteur près de la jonction et l'autre due à la modulation de conductivité.

Un modèle simplifié de la surface de la jonction et l'équation de diffusion sont utilisés pour déterminer l'excédent de densité de porteur en fonction du temps et de la position. En partant de la fonction de l'excédent de la densité de porteur, on peut calculer analytiquement les deux composantes transitoires.

La théorie explique d'une manière satisfaisante tous les phenomènes observés. La diode se comporte en circuit inducteur en présence d'un courant avant élevé, en capacité en présence de faibles courants, et en circuit oscillateur en présence de certaines valeurs intermédiaires du courant. Aussi ces équations sont confirmées avec une précision de 5 pour cent ou mieux par les données expérimentales des diodes alliées au germanium et au silicium.

Plusieurs applications sont suggérées telles que les méthodes de dimensionnement pour minimiser la tension transitoire avant ainsi que les mesures de longévité et de résistance "en vrac" déduites des observations transitoires.

Zusammenfassung—Das Verhalten von Übergangsdioden beim Durchgang von Stromimpulsen in der Vorwärtsrichtung wird analysiert. Die theoretischen Ergebnisse werden experimentell geprüft. Der Vorwärtsübergang der Diode lässt sich in eine Übergangsspannungs- und eine Widerstandskomponente zerlegen. Die erstere rührt von dem allmählichen Ansteigen der Dichte der Überschussträger in der Nähe des Übergangs her, die letztere von der Modulation der Leitfähigkeit.

Ein vereinfachtes Modell mit Flächenübergang und eine Diffusionsgleichung dienen zur Bestimmung der Dichte der Überschussträger als Funktion von Zeit und Ort. Aus dieser Funktion lassen sich die beiden Übergangskomponenten analytisch bestimmen.

Die Theorie erklärt alle Beobachtungen in zufriedenstellender Weise: Die Diode wirkt bei grossem Vorwärtsstrom induktiv, bei kleinem Strom kapazitiv und bei gewissen intermediären Strömen als Oszillator. Durch Versuche mit Germanium- und legierten Siliziumdioden wurden die Gleichungen mit einer Genauigkeit innerhalb von 5 Prozent experimentell bestätigt.

Als Beispiele für praktische Anwendung ergeben sich: Die Entwicklung von Methoden zur Reduktion des Vorwärtsübergangs und die Messung von Lebensdauer und Widerstand durch Beobachtung des Übergangs.

INTRODUCTION

WHEN the bias of a diode is switched from forward to reverse, the transient behavior is referred to as the reverse recovery or reverse transient. Similarly, the forward transient of the forward recovery is used to denote the diode behavior when the bias is switched from reverse or equilibrium state to forward state. The study of the transient behavior of semiconductor diodes has been the subject of intensive interest in the past few years by electron device and circuitry workers. It is well known that the diode behaves very differently from that predicted from its static behavior when the bias is changed abruptly, in either the forward or reverse directions. This transient behavior has been shown to be caused by the change of carrier densities, corresponding to different bias conditions, as well as by the conductivity modulations.[1-5]

It is the purpose of this article to examine in detail and to explain quantitatively the forward transient behavior of semiconductor junction diodes. A simplified but reasonable model is used to provide a unified explanation for the results reported under different conditions.

A number of workers have studied this problem, and this article also reports results of new studies. The important reported characteristics may be tabulated as follows:

(a) If a constant step voltage is applied to a diode in equilibrium and the steady-state current is quite large, the current as a function of time is as shown in Fig. 1(a).
(b) If a large step current is applied, the voltage function is as shown in Fig. 1(b).
(c) For small currents the shapes of the voltage and current curves are the transposed forms under (a) and (b).
(d) The initial condition of the diode, either in the reverse state or in the equilibrium state, has little effect on the observed forward transient behavior.
(e) If the diode is biased slightly in the forward direction initially, the maximum deviation between the initial value and the final value (voltage or current) can be reduced greatly.

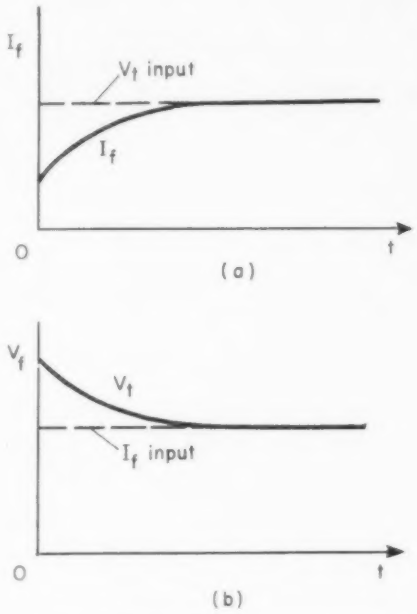


FIG. 1. The forward transient curves.

Measurements made by the author gave evidence that the diode forward transient consists of two components; one behaving in a manner

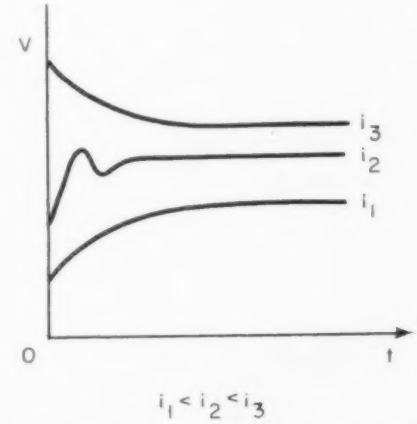


FIG. 2. Forward transient voltage characteristics.

similar to an inductive circuit, and the second in a manner similar to a capacitive circuit. For large currents the inductive character dominates; while for small currents, it is the capacitive behavior that dominates. Between the two extremes an oscillatory behavior was observed. Fig. 2 illustrates the three typical characteristics of forward transient voltage for a step current.

PHYSICAL MODEL AND QUALITATIVE EXPLANATION

For simplicity, the following assumptions are made which constitute the model of the junction diode to be studied.

(1) The diode can be represented by a planar diode as shown in Fig. 3.

(2) The conductivity of the p-region, σ_p, is much greater than that of the n-region, σ_n.

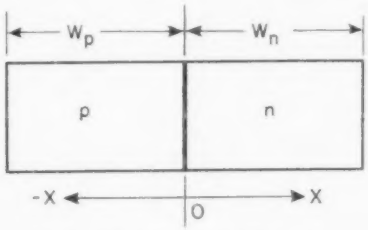


FIG. 3. The planar diode.

(3) The width of the n-region, W_n, is large compared with the diffusion length of holes in the n-region, L_p.

(4) The minority-carrier density at position X and time t, $p_n(X,t)$, is small compared with the majority-carrier density in equilibrium, n_{n0}.

(5) The transition region is narrow and abrupt.

(6) The metal contact at both terminals may be considered as ohmic contacts.

Furthermore, the following symbols are defined:

(a) $p_n(X,t)$ is the hole density in the n-region at time t and at a distance X from the junction.

(b) $p(X,t) = p_n(X,t) - p_{n0}$ is the injected hole density.

(c) $p(x,r)$ is obtained by substituting $x = X/L_p$, $r = t/\tau_p$, in $p(X,t)$, where x is the normalized distance, r the normalized time, L_p the diffusion length of holes in the n-region, and τ_p the lifetime of holes in the n-region.

(d) Capital letters V or I are used to represent steady-state quantities or some particular values of voltage or current, while v and i represent instantaneous or varying quantities.

A qualitative explanation of the forward transient phenomena may be obtained by considering the following physical description.

Under the assumptions made, the diode transient behavior is determined by the carrier densities in the n-region. The steady-state carrier densities in the diode at equilibrium as well as when biased to a forward current, are shown in Fig. 4. Where the abrupt junction is represented

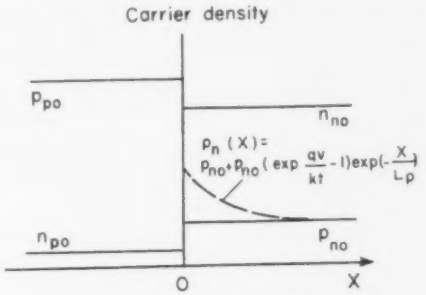


FIG. 4. Carrier densities of the diode.

by the line $X = 0$; p_{n0} and n_{n0} are the equilibrium carrier densities in the n-region; and $p_n(X)$ is the minority-carrier density when the diode is subject to a forward bias. This bias causes a current I_f to flow and establish a junction voltage V. The steady-state injected hole density is, as defined,

$$p_n(X) = p_{n0}\,[\exp(qV/kT) - 1]\exp(-X/L_p)$$

If, at time $t = 0$, a constant forward current is applied to the diode which is initially in the equilibrium state, the injected hole density has to build up from zero toward the steady-state value in a manner as shown in Fig. 5. Accordingly, the voltage across the junction v increases with time from zero toward the final value V. At the same time, the conductivity of the material is increased due to the increased majority- and minority-carrier densities. Therefore, the bulk resistance of the diode is reduced; this, in turn, causes the ohmic drop to decrease. The terminal voltage of the diode is equal to the sum of the junction voltage and the ohmic drop. The junction voltage increases with time, while the ohmic drop decreases with

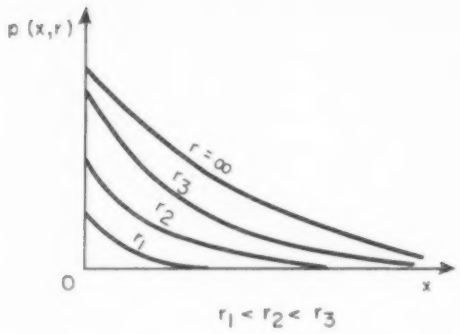


FIG. 5. The hole-density distributions in the forward transient period.

time. For very small current, the ohmic drop is small compared with the junction voltage. The observed terminal voltage increases with time. For very large currents, the junction voltage has reached its saturation value, and the ohmic drop is the dominant one. In this case, the terminal voltage decreases with time. Over some intermediate values of current the terminal voltage rises to a high value, then decreases to a minimum, and then increases again toward the steady-state value. Fig. 6 shows photographs of the forward-voltage transients observed.

MATHEMATICAL THEORY

In the n-region of the diode model described above, the injected hole density $p(x,r)$ can be determined from the diffusion equation

$$\frac{\partial p(x,r)}{\partial r} = \frac{\partial^2 p(x,r)}{\partial x^2} - p(x,r) \tag{1}$$

where $x = X/L_p$ and $r = t/\tau_p$. This equation is solved by means of Laplace transform subject to the following conditions:

initial condition $\quad p(x,0) = 0 \tag{2a}$

boundary condition $\quad p(\infty,r) = 0 \tag{2b}$

$$J_f = (I_f/A) = (-qD_p/L_p)\left(\frac{\partial p(x,r)}{\partial x}\right)_{x=0} \tag{2c}$$

$$= \text{constant}$$

final condition

$$-p(0,\infty) = p_{n0}(e^{aV}-1) = I_f L_p/qD_p \tag{2d}$$

where I_f is the forward current, A the junction area of the diode and $a = q/mkT$ where $m = 1$ for small I_f and $m = 2$ for larger I_f. The solution of equation (1) subject to the conditions given by equations (2) is

$$p(x,r) = p(0,\infty)\, s(x,r) \tag{3}$$

where

$$s(x,r) = \tfrac{1}{2}\left[e^{-x}\text{erfc}\left(\frac{x}{2\sqrt{r}} - \sqrt{r}\right) - e^{x}\text{erfc}\left(\frac{x}{2\sqrt{r}} - \sqrt{r}\right)\right] \tag{4}$$

By applying the boundary conditions, the transient voltage across the junction at normalized time r is

$$v(r) = \frac{kT}{q}\ln\left[1+(e^{aV}-1)\,\text{erf}\sqrt{r}\right] \tag{5}$$

The total injected minority carriers stored in the n-region at a time $t = \tau_p r$ is

$$Q_s(r) = qA\int_0^{W_n} p(X,t)dX = \int_0^{x_n} qAL_p(0,\infty)s(x,r)dx \tag{6}$$

where $x_n = W_n/L_p$. For $x_n \gg 1$,

$$Q_s(r) = \tau_p I_f(1-e^{-r}) \tag{7}$$

The conductivity in the n-region, when the diode is in thermal equilibrium, is $\sigma_0 = q\mu_p(p_{n0}+bn_{n0})$, where μ is the mobility of the carrier and $b = \mu_n/\mu_p$. Accordingly, the resistance of the diode is

$$R_0 = \frac{L_p}{\sigma_0 A}x_n \tag{8}$$

After the forward current flows for a time, $t = \tau_p r$, the hole density, as well as the electron density, is increased. The conductivity is now

$$\sigma(x,r) = \sigma_0\left[1+\frac{(1+b)p(0,\infty)}{p_{n0}+bn_{n0}}s(x,r)\right] \tag{9}$$

Then the bulk resistance becomes

$$R(r) = \int_0^{x_n}\frac{L_p dx}{A\sigma(x,r)} = R_0 - R_m(r) \tag{10}$$

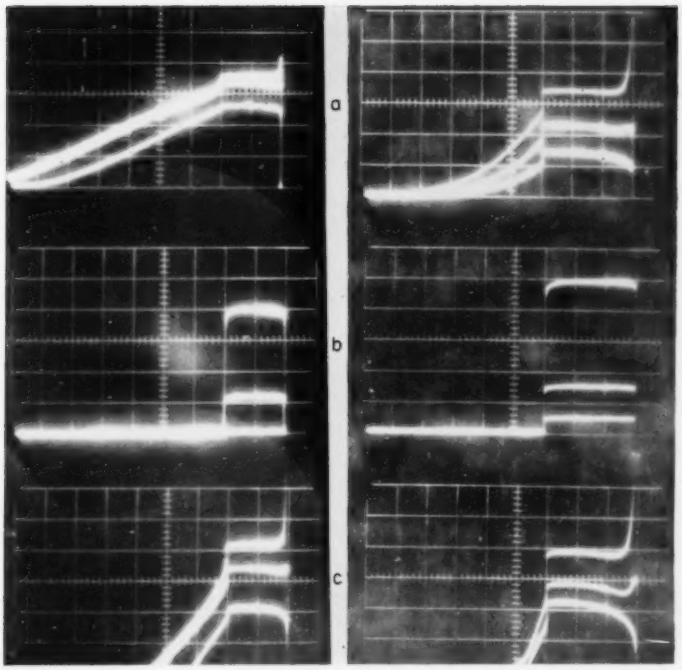


FIG. 6. The three types of forward transient voltage curves. (a) The terminal voltages; (b) the constant-current pulses; (c) the enlarged terminal voltages. The horizontal time axis is from right to left, its scale is 10 μsec/cm for Syl-III, and 50 μsec/cm for Clev-I. The vertical scales are: for Syl-III, (a) 0·2 V/cm, (b) 2·5 mA/cm, (c) 0·05 V/cm, for Clev-I, (a) 0·05 V/cm, (b) 0·5 mA/cm, (c) 0·02 V/cm.

[*facing p.* 62

where $R_m(r)$ represents the change in resistance due to the forward current. If we define

$$K = \frac{(1+b)\,p(0,\infty)}{p_{n0}+bn_{n0}} \tag{11}$$

Then

$$R_m(r) = \frac{L_p}{\sigma_0 A}\ln\left(\frac{1+K\,\mathrm{erf}\,\sqrt{r}}{1+Ks(x,r)}\right)$$

$$-\int_0^{x_n}\frac{2}{\sqrt{\pi}}\int_{[(x/2\sqrt{r})+\sqrt{r}]}^{\infty} e^{x-u^2}\,du\,dx \tag{12}$$

The second term in equation (12) yields an alternating series and makes a relatively small contribution to the total $R_m(r)$ when r is either very small or very large. For a first approximation

$$R_m(r) \simeq \frac{L_p}{\sigma_0 A}\ln\frac{1+K\,\mathrm{erf}\,\sqrt{r}}{1+Ks(x,r)} \tag{13}$$

Equation (13) may be derived also by using $s_1(x,r) = \mathrm{erf}\,(\sqrt{r})\exp(-x)$ to approximate $s(x,r)$ in equation (10). By substitution of equation (13) into equation (10), for $x_n \gg 1$,

$$R(r) \simeq R_0\left[1-\frac{1}{x_n}\ln(1+K\,\mathrm{erf}\,\sqrt{r})\right] \tag{14a}$$

while for $x_n > 1$

$$R(r) \simeq R_0\left[1-\frac{1}{x_n}\ln\left(\frac{1+K\,\mathrm{erf}\,\sqrt{r}}{1+K\,\mathrm{erf}\,\sqrt{r}\exp(-x_n)}\right)\right] \tag{14b}$$

When $r \gg 1$, equation (14a) becomes the same as the equations derived by BARNES[2] and by SHULMAN and MCMAHON[5].

The terminal voltage of the diode, v_t, is the sum of the junction voltage and the resistance drop. From the static diode equation $I_f = I_s(e^{aV}-1)$ and equation (5), and (14a) for $x_n \gg 1$,

$$V_t(t) = \frac{kT}{q}\ln\left(1+\frac{I_f}{I_s}\,\mathrm{erf}\sqrt{r}\right)$$

$$+I_fR_0\left[1-\frac{1}{x_n}\ln(1+K\,\mathrm{erf}\sqrt{r})\right] \tag{15}$$

At $t = 0$ $(r = 0)$, $v_t(0) = I_fR_0$; at $t \gg 1$ $(r \gg 1)$,

$$v_t(\infty) = V = \frac{kT}{q}\ln\left(1+\frac{I_f}{I_s}\right) + I_fR_0\left[1-\frac{1}{x_n}\ln(1+K)\right] \tag{16}$$

From equation (15) and equation (16), the transient voltage curves in Fig. 6 are readily explained. For $v_t(0) \gg v_t(\infty)$, the diode behaves inductively, while for $v_t(0) \ll v_t(\infty)$ it behaves capacitively. For $v_t(0) \simeq v_t(\infty)$ the diode behaves like a heavily damped oscillatory circuit.

The transition capacitance c_t of the diode is in parallel with the junction. When the junction voltage changes with time, the capacitance will require some charging current which will slow down the rate of change of transient voltage. However, for the forward transient of diodes, this effect is, generally, small and can be neglected. This is justified by comparing the capacitance current I_c, with the diode current I_f. For $x \gg 1$ from equation (5)

$$I_c = c_t\frac{dv}{dt}\,\frac{c_t}{\tau_p a}\,\frac{I_f}{I_s}\,\frac{e^{-r}}{\sqrt{(\pi r)}}\,\frac{1}{1+(I_f/I_s)\,\mathrm{erf}\sqrt{r}}$$

Therefore,

$$\frac{I_c}{I_f} = \frac{c_t}{\tau_p a}\,\frac{e^{-r}}{\sqrt{(\pi r)}}\,\frac{1}{(I_s+I_f\,\mathrm{erf}\sqrt{r})}$$

The maximum ratio occurs at $r = 0$, and decreases rapidly as r increases. Assume $I_f = 1$ mA, $I_s = 10\ \mu$A and $c_t = 10^{-6}\,\tau_p$ F. For $r > 0{\cdot}01$ the I_c/I_f ratio is less than 10^{-3}.

If the diode is not started from an equilibrium state but with an initial current I_1, then the amplitude of the resistance transient may be reduced. This results from the fact that in this case the initial resistance, instead of being R_0, is

$$R_1 = R_0\left[1-\frac{1}{x_n}\ln(1+K_1)\right]$$

where

$$K_1 = \frac{(1+b)p_{n0}}{p_{n0}+bn_{n0}}\,\frac{I_1}{I_s}$$

If I_1 is set equal to δI_f, then $K_1 = \delta K$. Therefore, the maximum change of resistance is now

$$R_1 - R(\infty) = \frac{R_0}{x_n}\ln\left(\frac{1+K}{1+\delta K}\right) \tag{17}$$

On the other hand, if the diode is started from the equilibrium state, the maximum change of resistance is

$$R_0 - R(\infty) = \frac{R_0}{x_n} \ln(1+K) \tag{18}$$

If we define the coefficient ρ_1 as the ratio of the maximum changes in equation (17) and equation (18),

$$\rho_1 = \frac{R_1 - R(\infty)}{R_0 - R(\infty)} = 1 - \frac{\ln(1+\delta K)}{\ln(1+K)} \tag{19}$$

The reduction of ρ_1 by an initial current δI_f can be related to the transient resistance curve in the following way. Consider the ratio of maximum resistance change from $r = r_1$ to $r = \infty$ to the maximum change from $r = 0$ to $r = \infty$, and call it ρ_r. Then

$$\rho_r = 1 - \frac{\ln(1+\text{erf}\sqrt{r_1 K})}{\ln(1+K)} \tag{20}$$

Consequently, if $\text{erf}\sqrt{r_1} = \delta$, then $\rho_1 = \rho_r$. This implies that the reduction in the resistance transient due to an initial bias current $I_1 = \delta I_f$ is equivalent to starting the transient at $t_1 = r_1 \tau_p$, where $\text{erf}\sqrt{r_1} = \delta$.

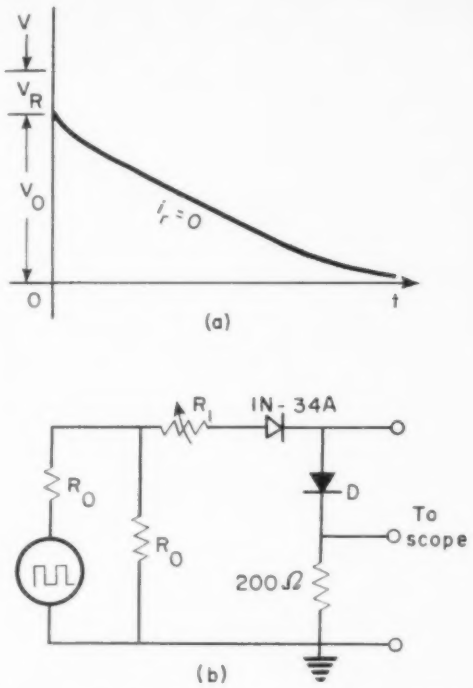


FIG. 7. Open-circuit diode voltage curves and the test circuit.

EXPERIMENTAL RESULTS

The technique developed by GOSSICK[4] for separating the junction voltage from the ohmic drop was used in the following experiments. If a diode in a forward state is suddenly open-circuited, the voltage across the diode is shown in Fig. 7(a). The initial jump V_R, is the ohmic drop of the diode and V_0 is the junction voltage at $t = 0$.

The test circuit used is shown in Fig. 7(b). R_1 was used to adjust the forward current. The IN-34-A diode provided an open-circuit condition after the forward-current pulse was over. The recovery time of IN-34-A was selected to be shorter than 0·05 μsec. The current was measured from the voltage drop across the 200-Ω precision resistance. The pulse repetition rate was adjustable from 50 to 200 pulses/sec, the time interval between two pulses was kept greater than twenty times the measured lifetime of diode D. The possible error in measurements is 5 mV. Data

Table. 1 *The properties of sample diodes*

No.	Description	Wafer area (mil²)	W_n (mil)	x_n	Resistivity (Ω-cm)	R_0 (Ω.)
Clev-1	n-Ge, alloyed	80 × 80	35*	1·4	20–25	138
GE-III	n-Ge, alloyed	45 cir. mil	11	1·0	20–25	
Syl-III	n-Si, alloyed	47 × 47	16	4	15–25	70
HD-6006	n-Si, alloyed	32 × 32	19	5		88

* Measured value from a similar sample.

taken with special care may be accurate within 2·5 mV.

The physical dimensions as well as the resistivities of the sample diodes are listed in Table 1. It is known that in the diode equation

$$I = I_s(e^{aV}-1)$$

the value a equals q/mkT, for small currents $m = 1$; for large currents $m = 2$.[6] For the silicon samples and the germanium samples with medium currents, the quantity a from the static characteristics, has the values shown in Table 2.

As a verification of equation (5), refer to Fig. 8(a) and (b) where the solid theoretical curve is calculated from the measured steady-state-junction voltage V, while the data points are represented by x. For small voltages, [curves 1 and 2 in Fig. 8(a)], $a = q/kT = 39$ was used; for larger voltages, [curves 3 and 4 in Fig. 8(a) and curves in Fig. 8(b)], the measured values of a were used. Though the theory is derived from the low injection assumption, it is found that the results can be applied to the large-current cases, provided the actual value of a is used and the change of lifetime during this transient period be corrected (see Appendix).

Table 2. The values of "a" of sample diodes

Sample diode	Germanium 1 Clev-1	Silicon 1 Syl-III	Silicon 2 HD-6006
a (V^{-1})	27	23	28
At I (mA)	2	0·7	5

For small currents, data from the germanium diode Clev-1 and the silicon diode Syl-III showed a maximum error of 2·5 mV as compared with the theoretical values. For large currents, the maximum error was 5 mV. They are all within the expected range of experimental error.

In order to test the equations derived for transient resistances, R_0 was measured by a very short current pulse (0·1 μsec), $R(\infty)$ by a very long current pulse ($10\tau_p$), and K calculated by equation (14) with $r = \infty$. By substitution of these values in equation (14a) or (14b) the theoretical resistance transient curve may be

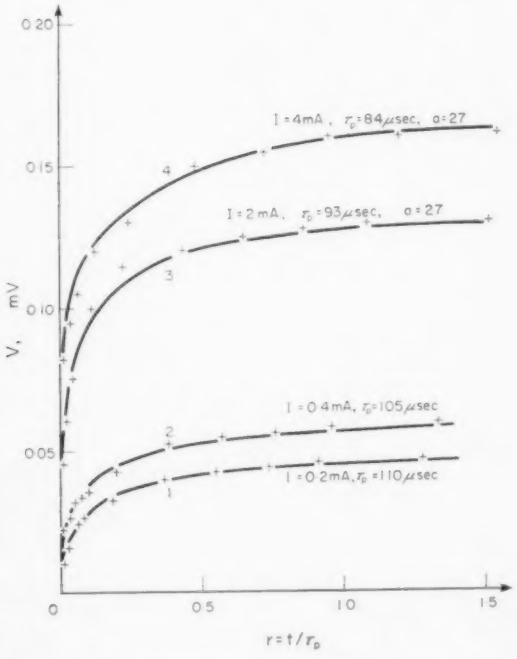


FIG. 8(a). The forward-voltage transient characteristics of a germanium diode (Clev-I).

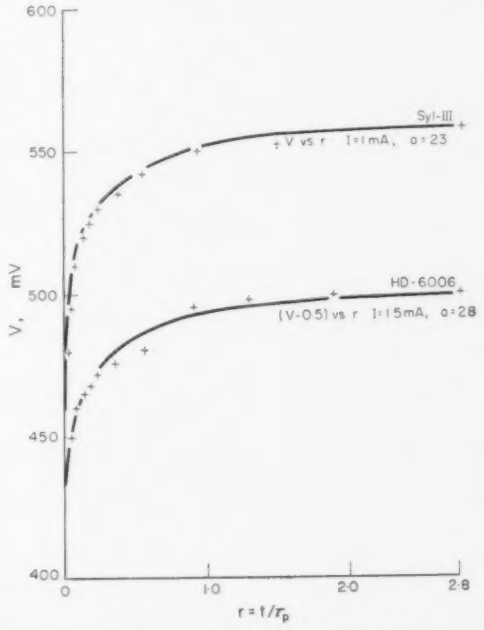


FIG. 8(b). The forward-voltage transient characteristics of silicon diodes (with zero suppressed ordinate).

calculated. The comparison between the actual data and the theoretical curve is shown in Fig. 9 (a) and (b). The agreement is good. Just as expected, the approximation equation (14), gives better accuracy at small and large values of r than at intermediate ones.

The diode terminal transient voltage, $V_t(r)$, of

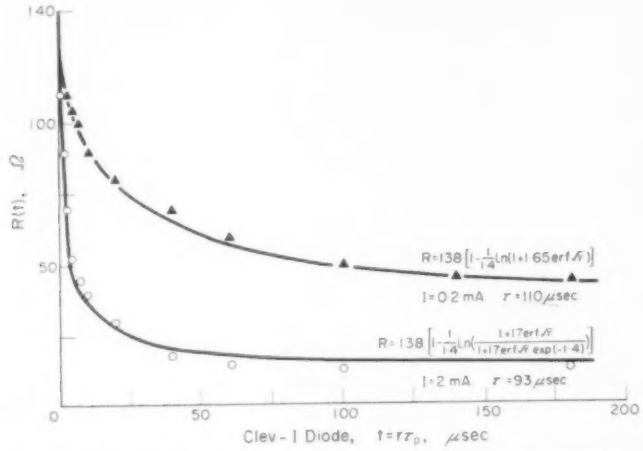


FIG. 9(a). The forward-resistance transient characteristic of a germanium diode.

Clev-1 sample is calculated from equations (5) and (14b) at $I_f = 0{\cdot}2$ mA, $0{\cdot}8$ mA and 2mA, and is plotted in solid lines in Fig. 10. The data points in the same figure represent the measured values.

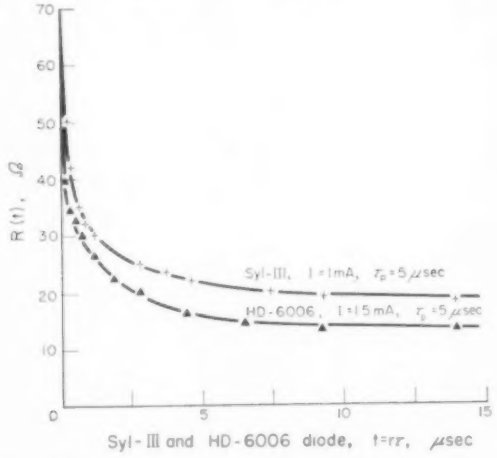


FIG. 9(b). The forward-resistance transient characteristics of silicon diodes.

The agreement between the analytical calculation and the experimental measurement is good for all three types of transient voltage.

The reduction of the inductive transient due to an initial current $I_1 = \delta I_f$ is equal to starting the transient at $r_1 = t/\tau_p$ with $\text{erf}\sqrt{r_1} = \delta$. In Fig. 9, it is clearly illustrated that the larger the current the sharper the decrease of the transient resistance. Therefore, for a constant δ, the larger the final current, the larger the reduction of the transient, and the smaller the value of ρ_1. This is illustrated by the three curves shown in Fig. 11(a). Fig. 11(b) shows the effects of ρ_1 if x_n is not infinity. It is observed that the smaller the value of x_n the more rapidly ρ_1 decreases with δ.

CONCLUSIONS

From the study here reported the following conclusions are possible:

(a) The forward transient consists of two components: the junction-voltage transient and the diode-resistance transient. The former increases with time and the latter decreases with time. These transients may be calculated accurately from equations (5), (12) and (15).

(b) Due to the different functional relationships between the two transient components with the forward current, three types of over-all transients are possible, as illustrated in Figs. 6 and 10.

(c) For minimum overshoot and fast rise of the

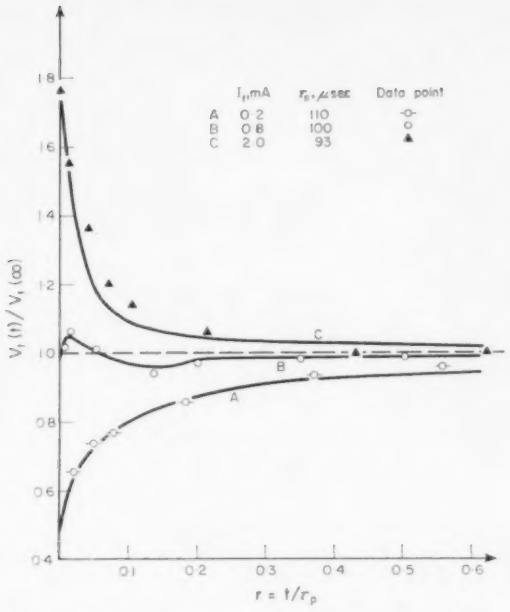


FIG. 10. The diode-terminal transient voltages of a germanium diode (Clev-I).

terminal transient voltage, the initial voltage should be nearly the same as the final one. For a given diode with known initial resistance R_0, and the static-terminal I–V characteristic, the optimum operating point A may be found as shown in Fig. 12.

(d) For large currents, the over-all transient of the diode behaves somewhat similarly to an RL circuit, the initial voltage being much greater than the steady-state voltage. The amplitude of the transient may be reduced by biasing the diode in the forward direction initially. The amount of reduction in transient amplitude is determined by the final current and the ratio of the initial current to the final current, as illustrated in Fig. 11. The approximate value of reduction can be calculated from equation (19) or from the resistance transient curve and equation (20).

(e) If the forward transient is to be eliminated the following methods may be used :

(i) Decrease the widths, W_n and W_p, of the end materials.

(ii) Use diode materials with smaller minority-carrier lifetime.

(iii) Use diode materials with large conductivity.

(iv) Design the circuit as described in (c) to give optimum performance.

(v) Bias the diode in the forward direction initially.

(f) Several applications may be suggested as follows.

(i) The forward-voltage transient curve [equation (5)] is shown to have an accuracy of 5 per cent or better; it can be used to measure the lifetime of the minority carrier in the diode material[7] for $[qv(kT)] \gg 1$ where v is the instantaneous junction voltage.

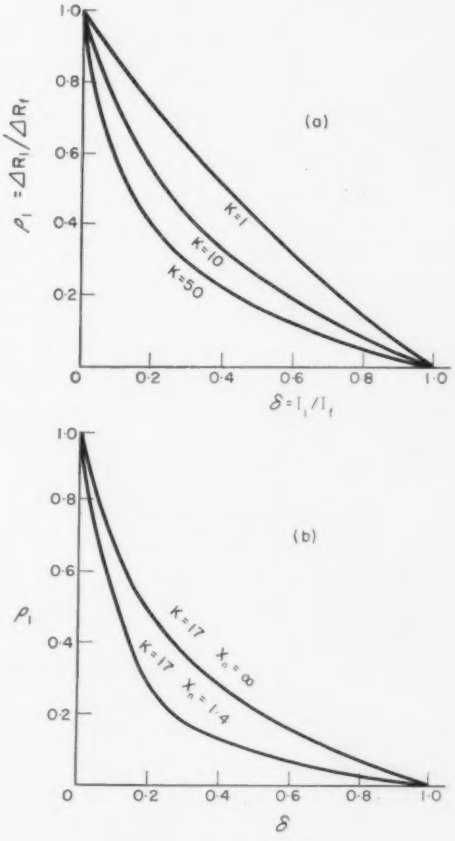


FIG. 11. ρ_1 vs. δ curves.

(ii) For the assumed model, the junction injection ratio is unity. The charge stored in

the diode end material by a current I_f for a time t is, from equation (7),

$$Q_s = \tau_p I_f[1-\exp(-t/\tau_p)]$$

If the injection ratio is not unity but γ, then

$$Q_s = \gamma\tau_p I_f[1-\exp(-t/\tau_p)]$$

This may be used to measure the injection ratio. Q_s can be determined by reverse recovery method.[7, 8]

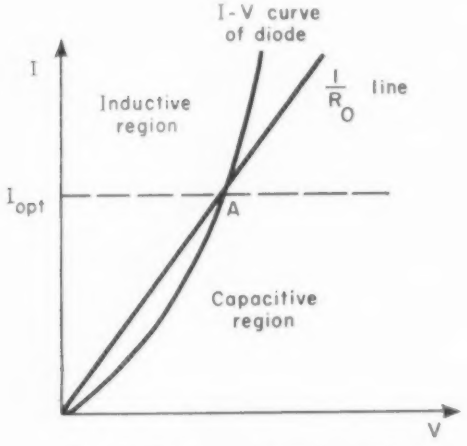


FIG. 12. The forward transient regions.

(iii) A method to measure the diode bulk resistance is provided by the pulse technique as discussed.[7]

(iv) The static characteristic may be expressed as

$$I = I_s[\exp\{a(V-IR)\}-1]$$

$$R = R_0\left[1-\frac{1}{x_n}\ln\left(1+\frac{(1+b)\,P_{n0}\,I}{P_{n0}+bn_{n0}}\,\frac{I}{I_s}\right)\right]$$

(v) A means to measure I_s may be developed from the accurate static characteristics.

Acknowledgements—The author wishes to acknowledge the encouragement and assistance from Dr. F. E. BRAMMER and Dr. S. SEELY of Case Institute of Technology, and Dr. A. M. GOODMAN of RCA Research Laboratory at Princeton, New Jersey.

REFERENCES

1. C. DORN, *I.R.E. Trans.* **ED-3,** 153 (1956).
2. F. BARNES, *Proc. I.R.E.* **46,** 1427 (1958).
3. H. ARMSTRONG, *I.R.E. Trans.* **ED-4,** 111 (1957).
4. B. GOSSICK, *J. Appl. Phys.* **26,** 1356 (1955); **27,** 905 (1956).
5. R. SHULMAN and M. MCMAHON, *J. Appl. Phys.* **24,** 1267 (1953).
6. M. CULTER, *J. Appl. Phys.* **26,** 949 (1955).
7. W. KO, Thesis, Case Institute of Technology (1959).
8. O. CURTIS and B. GOSSICK, *Rev. Sci. Instrum.* **27,** 828 (1956).
9. G. BEMSKI, *Proc. I.R.E.* **46,** 990 (1958).
10. W. SHOCKLEY and W. READ, *Phys. Rev.* **89,** 835 (1952).

APPENDIX

The minority-carrier lifetime in the diode material has been derived as a function of the injected carrier density Δp, as[9, 10]

$$\tau = \tau_0\frac{1+a'\Delta p}{1+b'\Delta p}$$

Where a' and b' are constants, and τ_0 is the low-level lifetime. Because the injected minority-carrier density is proportional to the forward current, it is obvious that τ will vary with forward current in the same manner that it varies with Δp. The measured lifetimes from reverse

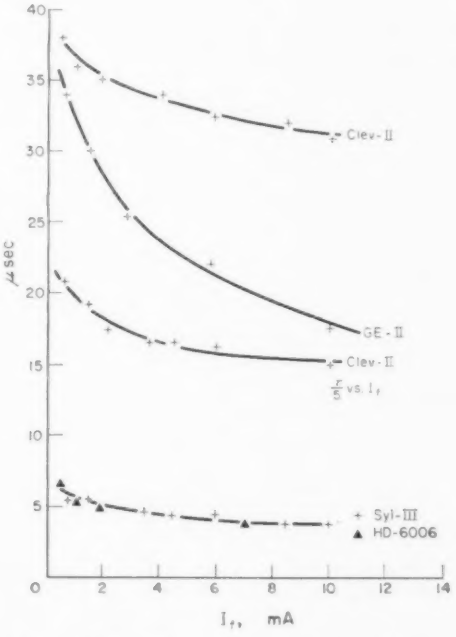


FIG. 13. The lifetime vs. forward-current curves.

recovery time are plotted against current in Fig. 13. These curves show good agreement with the above equation. During different periods of the current flow, according to equation (3), the injected hole density varies from zero to the final value. It is expected that the lifetime will change during the forward transient period. Therefore, in the calculation of $r = t/\tau_p$, the proper values of τ_p corresponding to the hole density should be used. In the calculation of curves 1 and 2 in Fig. 8(a), uniform τ_p is used because the error without correction is smaller than experimental error. For curves 3 and 4 in Fig. 8(a) and curves in Fig. 8(b) the change of τ_p cannot be neglected; a linear change of τ_p is assumed to give a first-order correction. The correction is appreciable (2–5 per cent) and it improves the agreement of experimental data with the calculated curve.

Following the same reasoning as discussed for lifetime, the value of $a = q/mkT$ actually does change with time in the transient period. For more accurate results the value of a should be modified.

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 70–75. Printed in Great Britain

NOTES

Ge or Si junctions made by the surface weld-regrowth process

(*Received* 2 *February* 1961; *in revised form* 13 *March* 1961)

A TECHNIQUE for making p–n junction structures of germanium or silicon, which consists of the melting, wetting and resolidification of one piece of semiconductor on the surface of another, unmelted piece, has been described by LESK[1]. A similar process, which produces junctions of the same metallographic appearance as those described by LESK, was developed independently in this laboratory. Since the LESK paper does not present data on the electrical characteristics of devices made by his process, it is the purpose of this note to present some characteristics of p–n junctions made in our work, and to show that they compare favorably to junctions made by other means. Because of the nature of the process and the metallographic appearance of junction structures made by it, we call this process surface weld–regrowth.

Fig. 1 shows the direct-current characteristics, measured at room temperature, of four p–n junctions made of low-resistivity germanium. Using a resistance-heated thermal-gradient furnace, arsenic-doped germanium of 7×10^{-4} Ω-cm resistivity was melted and regrown on the surface of indium-doped germanium of 0·25 Ω-cm resistivity to form junction areas approximately circular in shape, and measuring from 0·030 to 0·035 in. in diameter. Ohmic contacts were made to the n- and p-type germanium with antimony–gold alloys and lead–indium alloys, respectively, attached at temperatures below 400°C. The reverse characteristics were developed by using either immersion etching in CP-4, or electrolytic etching in a potassium hydroxide solution. Similar results were obtained from the two etches. It may be noted that the junctions do not show a rigid saturation of the reverse current. The same effect may be noted in germanium junctions made by a vapor growth process.[2]

Fig. 2 shows characteristics of three junctions made with antimony-doped germanium of 0·4 Ω-cm resistivity as the melted and regrown material. The substrate was indium-doped germanium, of 0·25 Ω-cm resistivity for one of the

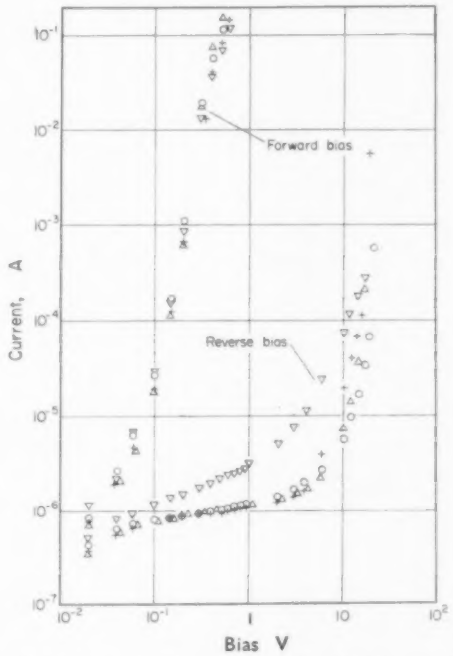


FIG. 1. Direct-current characteristics of four germanium junctions made by surface weld–regrowth. 7×10^{-4} Ω-cm As-doped germanium regrown on 0·25 Ω-cm In-doped germanium.

units, and 4 Ω-cm resistivity for the other two. Other details of the processing were the same as for the units of Fig. 1. It may be noted that the saturation of the reverse current is more pronounced for the units of Fig. 2, although the currents are about ten times as great. Since the junctions represented in Fig. 2 are comparable in size (0·022–0·037 in. diameter) to those of Fig. 1, and since one of them uses the 0·25 Ω-cm substrate, we attribute the larger currents to the high-

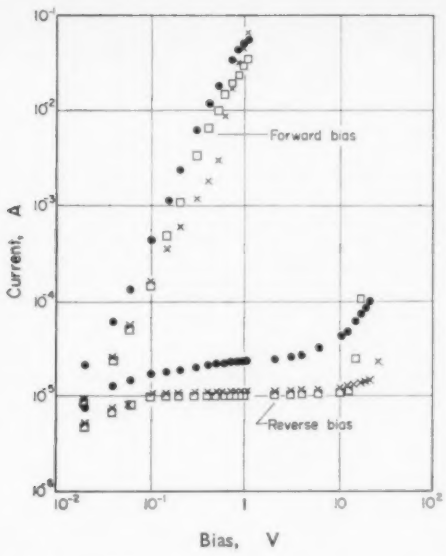


FIG. 2. Direct-current characteristics of germanium p–n junctions made by surface weld–regrowth. ×—Sb-doped germanium, 0·4 Ω-cm, regrown in In-doped germanium, 0·25 Ω-cm; ●, □—the same melt material, regrown on In-doped germanium, 4 Ω-cm.

resistivity regrown material produced by the antimony-doped germanium, and to low minority-carrier lifetimes caused by the severe heating of the surface weld–regrowth process. Consideration of the resistivities before melting, as well as the distribution coefficients of arsenic and antimony[3] would predict a much higher resistivity in the units of Fig. 2. Since resolidification proceeds from the junction outward due to the thermal gradients used to achieve preferential melting, a dopant of small distribution coefficient tends to leave the first regrown layers very lightly doped. The units of Fig. 2 also have poorer forward characteristics than those of Fig. 1. The swerve in the high-current portion of the forward characteristic of one (denoted by the symbol ×) is due to an unstable ohmic contact.

These junctions are compared to a theoretical model and to junctions made by other means in Fig. 3. The theoretical model is derived from the familiar equation[4]

$$I = I_s(e^{qV/kT} - 1) \qquad (1)$$

where I = the net current through the junction, I_s = the saturation current, q = the electronic charge, V = the applied bias, k = Boltzmann's constant and T = the absolute temperature. For any value of applied bias, the model predicts a ratio of forward to reverse current of

$$\frac{I_f}{I_r} = -e^{qV/kT} \qquad (2)$$

where the subscripts f and r, respectively, denote forward- and reverse-bias conditions. A plot of this function is given in Fig. 3, together with plots of

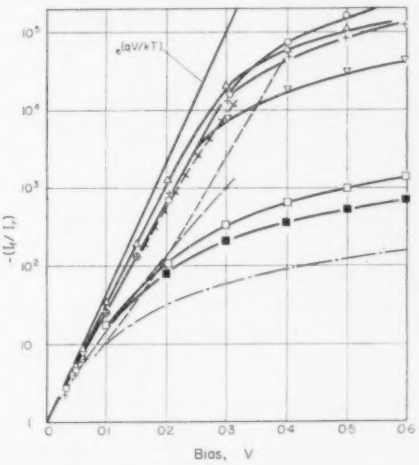


FIG. 3. Comparison of the rectification properties of germanium p–n junctions made by various means. ○△▽+—Surface weld–regrowth of As-doped, 7×10^{-4} Ω-cm on In-doped, 0·25 Ω-cm; □ ■—Surface weld–regrowth of Sb-doped 0·5 Ω-cm on In-doped, 4 Ω-cm.

— — — Vapor-grown P-doped, followed by vapor-grown B-doped

–+– –+– Same as above, but more heavily doped

– ——— – Vapor grown Ga-doped on Sb-doped, melt-grown substrate

—— —— Germanium junction, commercial, alloyed

experimentally determined values of (I_f/I_r) for junctions made by the surface weld–regrowth process, and others. Data for the other units were obtained from Ref. 2. Since the range of bias in Fig. 3 is below the reverse-bias breakdown of any of the samples, the sharp veering of the plots away from the theoretical model is due to the effect of resistance in series with the junctions on their forward characteristics. It is apparent that germanium

junctions made by surface weld-regrowth techniques have electrical characteristics which compare favorably with those of germanium junctions made by other means. This is particularly the case for the regrowth of heavily arsenic-doped germanium on low-resistivity indium-doped substrates. It is possible that their characteristics might be improved by heat treatments or alloy gettering,[1] although none of the germanium samples were given these treatments.

A less extensive study has been made of *p–n* junctions made by surface weld-regrowth of silicon. The samples were made by the regrowth of silicon doped heavily with boron, on arsenic-doped, 2 Ω-cm resistivity substrates. The samples were annealed for 13 hr at 475°C in argon in order to restore minority-carrier lifetimes. Reverse characteristics were developed by immersion etching in CP-4 (without bromine). Table 1 summarizes the low-voltage electrical characteristics of four units at room temperature.

Table 1. *Summary of direct-current characteristics, silicon junctions*

Sample no.	V_f at $I_f = 100$ mA (V)	I_f/I_r at $V_f = V_r = 0{\cdot}7$ V	I_r at $V_r = 1{\cdot}0$ V (μA)
4	0·97	$4{\cdot}8 \times 10^5$	0·165
5	0·96	$2{\cdot}8 \times 10^5$	0·230
16	1·10	$4{\cdot}5 \times 10^6$	0·006
20	0·81	$4{\cdot}0 \times 10^6$	0·018

The characteristics given above for germanium and silicon junctions made by the surface weld–regrowth technique affirm the usefulness of this and similar processes for making experimental semiconductor device structures.

T. C. Taylor
F. L. Yuan

Semiconductor Division
Raytheon Company
Newton, Mass.

References

1. I. A. Lesk, *J. Electrochem. Soc.* **107,** 534 (1960).
2. M. J. O'Rourke, J. C. Marinace, R. L. Anderson and W. H. White, *IBM J. Res. Dev.* **4,** 256 (1960).
3. F. A. Trumbore, *Bell Syst. Tech. J.* **39,** 207 (1960).
4. W. Shockley, *Electrons and Holes in Semiconductors* p. 91. Van Nostrand, New York (1950).

Melting phenomena with silicon

(*Received* 7 *February* 1961; *in revised form* 22 *March* 1961)

Pearson and Treuting[1] have noted that small, randomly distributed regions on freshly prepared silicon surfaces appear to melt at temperatures several degrees below the melting point of pure silicon. Some observations made in connexion with the development of the regrown–diffused transistor[2] may throw a little more light on this behaviour.

Blocks of freshly etched silicon* with (111) oriented main face were partially melted in a steep temperature gradient under pure dry hydrogen, the experimental arrangement being illustrated in Fig. 1. Before heating, such blocks showed no distinguishable surface features even at × 1000.

On examining the surface of the block close to the interface between melted and unmelted regions, characteristic striae would be observed. Fig. 2 shows a typical specimen: the darker area at the top is the melted region, and the striae appear to be the tracks of small globules of molten material on the unmelted silicon which have migrated up the temperature gradient, and which originated in a region showing numerous small black "nuclei".

* Degreased, boiled in conc. HNO_3, etched 1 min in 2 parts conc. HNO_3, 1 part 48 per cent HF, quenched in a large excess of double-distilled water, stream washed in a recirculator for 1 hr and dried under vacuum. CP-4 etch gave almost identical results, in spite of the "cobble-stone" surface obtained with it.

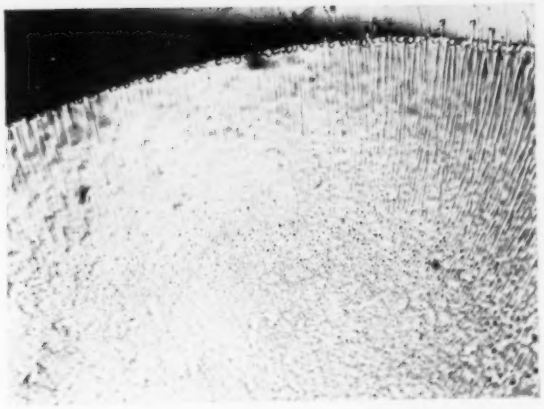


FIG. 2. Surface of silicon block near melt interface, after re-freezing, showing striae and "nuclei", ×66.

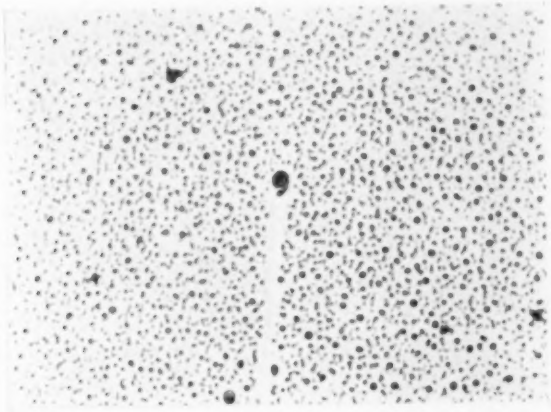


FIG. 3. Surface of silicon block near melt interface, after re-freezing, showing "nuclei", ×296.

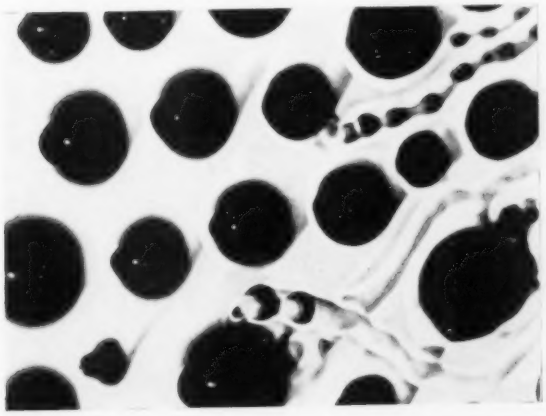


FIG. 4. Frozen silicon surface showing conical spikes, ×280.

[*facing p.* 72

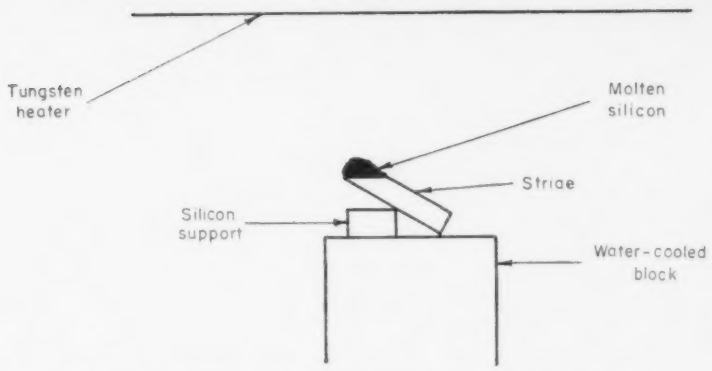


FIG. 1. Experimental arrangement for melting silicon.

This region, at higher magnification, is illustrated in Fig. 3 which also shows a globule in the process of formation. It should be noted that such globules freeze to form a typical conical spike, i.e. they are basically silicon.

That the globules actually migrate over the surface of solid silicon and not on a molten surface skin was proved by high-temperature observations using a long-working-distance microscope. The solid–liquid interface could clearly be followed during freezing as could the tracks of small globules climbing up the *solidified* silicon surface slightly below it. (The rate of climb could be relatively rapid, of the order of 0·5 mm/min.) Such tracks can also be seen "frozen" in Fig. 2.

It is suggested that impurities left on the surface depress the freezing point of silicon locally to form a small molten region and that this region migrates up the temperature gradient (possibly aided by surface-tension effects) by the temperature-gradient zone-melting mechanism postulated by PFANN[3]. An electron-diffraction examination of a surface showing "nuclei" indicated the presence of an unidentified material which was not silicon, silica or silicon carbide.

High-temperature observations also showed that, as a silicon surface melts, an extremely thin film of some material is drawn by surface tension to the edges and in particular to the corners of the molten region. This film is broken by numerous rounded holes (revealing the true silicon surface) which are in a state of constant agitation so long as melting proceeds. If power is reduced, a conical spike of silicon grows from each such hole, as shown in Fig. 4 (the film appears able to withstand the associated pressure). The black regions are the spikes. The spikes usually took up the same orientation as the unmelted material, as judged from the ⟨111⟩ ridges seen on all of them.

It was noted that during melting the thin surface film gradually tended to disappear, but whether this was due to evaporation or solution was not determined.

It is suggested that the film present on a molten surface and the "nuclei" on the solid have a common origin, and may be associated with a deposit left from the washing water. On evaporation this water left a very slight residue which was believed to come from the de-ionizing column used for purification.

C. H. L. GOODMAN*

The General Electric Company Limited
Hirst Research Centre
Wembley, England

Addendum

The melted regions observed by PEARSON and TREUTING showed outlines characteristic of the crystal face upon which they formed [i.e. triangular on a (111) surface]. No such orientation-dependent structure would be expected under the experimental conditions of the present work on account of the tendency for globules to move up the temperature gradient when they have formed.

In Fig. 2 the boundary between melted and

**Now at*: Standard Telecommunication Laboratories, Harlow, England.

unmelted regions is a "high-water mark", being the furthest point to which the liquid–solid interface penetrated. The tracks of globules crossing into the melted region formed after the interface had withdrawn, as briefly noted in the text.

The author is indebted to the referee for raising queries on these points.

Note added in proof

HANEMAN[4] has recently observed small molten regions moving on InSb at temperatures below the melting point. PEARSON and TREUTING referred to surface melting with such compounds and it thus seems probable that the effects described in the present Note will be encountered with other diamond-type materials, and possibly with pure congruently melting phases in general.

References

1. G. L. PEARSON and R. G. TREUTING, *Acta Cryst.* **11**, 397 (1958).
2. C. H. L. GOODMAN, *Solid-State Electron.* **1**, 188 (1960).
3. W. G. PFANN, *J. Metals, N.Y.* **7**, 961 (1955).
4. D. HANEMAN, *J. Appl. Phys.* **31**, 217 (1960).

Thin-film capacitors using tantalum oxide dielectrics prepared by reactive sputtering*

(*Received* 17 *February* 1961)

A METHOD of making small capacitor elements of the thin-film type is described and some results obtained from measurement of their electrical characteristics are given.

The form of these capacitors can be summarized as follows. A thin film of tantalum oxide is interposed between two aluminium-film electrodes, the whole structure being supported by a suitable substrate. The aluminium-film electrodes comprise the plates of the capacitor and the tantalum oxide its dielectric.

Details of the process for the formation of the capacitor elements are now given. A fused-quartz microscope slide, size $25 \times 76 \times 1$ mm, is prepared by washing in detergent, rinsing in distilled water and then isopropyl alcohol, and finally drying off by exposure to isopropyl alcohol vapour. This quartz substrate is then placed in a vacuum chamber where it is further cleaned by ionic bombardment preparatory to its coating with six base electrodes (evaporated from 99·99 per cent pure Al), their thickness being approximately 2000 Å. The substrate now carrying the base electrodes is then transferred to a sputtering chamber, similar in design to that described in the literature,[1] where it is coated with tantalum oxide by the chemical or "reactive" sputtering process.[2] In this process the oxide is formed by sputtering from a tantalum cathode in an oxidizing atmosphere provided by leaking oxygen into the vacuum system through a needle valve. The cathode potential was 2500 V and the cathode current density was maintained at 0·5 mA/cm², resulting in a sputtering pressure of 12μ Hg. These conditions gave a deposition rate of 350 Å/hr, limited by the temperature rise of the substrate. The thickness of the dielectric films formed under these conditions ranged from 1200 to 1800 Å. Finally the substrate is returned to the vacuum coating chamber where two aluminium counter electrodes are evaporated on to the surface of the tantalum oxide film, so forming twelve capacitor elements.

A diagrammatic representation of a set of twelve capacitor elements made as described above is shown in Fig. 1.

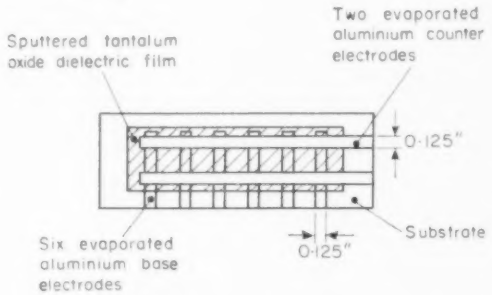


FIG. 1. Construction of twelve capacitor elements.

Capacitors formed in this way, for test purposes, have given capacitance values of 0·1 μF/cm², and have shown tan δ values of less than 0·01 up to 30,000 c/s. Frequency characteristics have been plotted for dissipation factor and capacitance, and are given in Fig. 2.

Unprotected films have withstood d.c. breakdown voltages up to 14 for a 1800 Å film; a

* U.K. Patent application No. 42094/60.

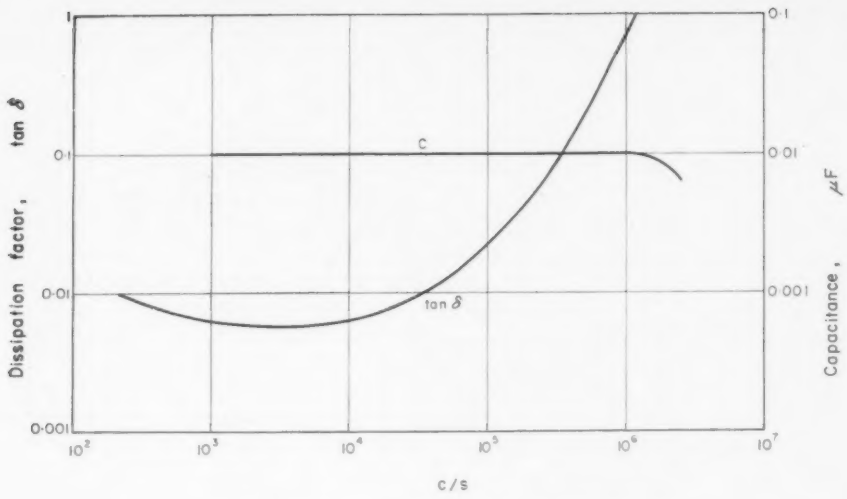


FIG. 2. Frequency characteristics of capacitor element.

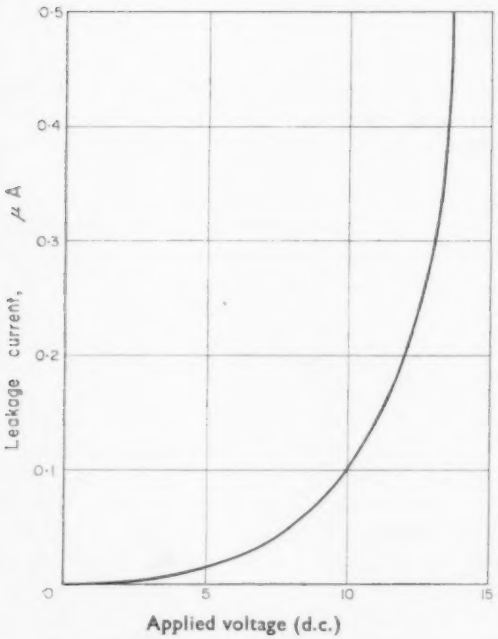


FIG. 3. Leakage current of capacitor element.

curve of leakage current against voltage obtained in this instance is shown in Fig. 3.

So far, except for periods of measurement extending to some hours, substrates have been stored in a desiccator; however, a small 300 kc/s oscillator has recently been constructed, using a single tantalum oxide capacitor element, which has been working satisfactorily under varied laboratory conditions without protection from moisture.

These reactively sputtered tantalum oxide film dielectrics should find particular application in the transistor field, either in individual capacitors or as component parts of integrated micro-circuit structures of the thin-film type. They would also appear to be of general interest where thin insulating inorganic films are required.

Royal Radar Establishment P. LLOYD
Great Malvern, Worcestershire

References

1. R. B. BELSER and W. H. HICKLIN, *Rev. Sci. Instrum.* **27,** 293 (1956).
2. L. HOLLAND and G. SIDDALL, *Vacuum* **3,** 245 (1953).

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 76–78. Printed in Great Britain

BOOK REVIEWS

RICHARD H. BUBE: **Photoconductivity of Solids.** Wiley, New York, 1960. 461 pp., $14.75.

Photoconductivity of Solids is literally in a class by itself. Books on photoconductivity are rare indeed, and there are no others available that attempt to deal with the whole subject. Thus the reader's choice is simple. If he wants a book that covers the whole field, this is it. Since it is the only book of its kind, one can only compare it with what might have been written and then try to assess the measure of success. The main functions of a book in an active field are to collect and summarize the available information, and, if possible, to digest and compact it. This latter step then provides a new base for further advances. *Photoconductivity of Solids* does an excellent job in the collecting and summarizing area. It has a good introduction to the present theoretical understanding, and presents a wide variety of experimental observations, including many from related fields. It may be less successful at building a new base from which to proceed, but that was really too much to hope for in the present state of the subject. Measured against a reasonable expectation, *Photoconductivity of Solids* is a good book, and is certainly the most convenient place to start the study of photoconductivity.

R. W. REDINGTON

E. H. PUTLEY: **The Hall Effect and Related Phenomena.** Butterworth, London, 1960. 263 pp., 50*s*.

THE title of this book is not an apt one, since the book is more nearly a review of the properties of semiconductors, including thermoelectric properties, magnetoresistance, thermal conductivity, and basic physical and chemical properties, than a book dealing with the Hall effect and its consequences. Thus the book is to be compared with similar books by the author of this review, by SPENKE, SMITH, and others. However, the title is justified in the sense that conduction properties are emphasized in the book, with little or no attention being given to optical properties, magnetic properties, and some other important aspects often covered in general reviews, such as crystal-growth phenomena, surface properties, etc. No attention is given to metals or dielectrics.

In the first section, "Conduction Processes", a brief general review is given of the elementary theory of conductivity and Hall effect, together with the basic model of the conduction due to impurities in semiconductors originally due to WILSON.

In the second section a valuable review is given of the methods of measurement of Hall effect and resistivity of semiconductors under various conditions, such as high and low temperatures. Various circuits for d.c. and a.c. measurements are shown, and various configurations for samples and sample holders are presented. This section is one of the most useful in the book since it is more extensive than can be found in most other reviews except for that in *Methods of Experimental Physics*, Vol. 6, Academic Press, New York (1959).

In Section 3, titled "Conduction in an Isotropic Solid", a review at a fairly high level is given of the Fermi–Dirac statistics, of the calculation of electrical conductivity, and of the general equations for the transport properties in the presence of a magnetic field. This material is less valuable than some of the other parts of the book since it is available in many other places, particularly in SMITH's book on semiconductors published in 1959.

In Section 4, "The Behaviour of Typical Semiconductors", a review is given of a large number of properties of semiconductors, including such effects as intrinsic and extrinsic conduction, mixed conduction by electrons and holes, impurity scattering phenomena, magnetoresistance phenomena, surface effects on galvanomagnetic properties, and the Nernst, Righi–Leduc, and Ettingshausen effects.

In Section 5, "A Summary of the Electrical Properties", a review is given of the conduction

properties of the elemental semiconductors of the germanium type, of the III–V compounds, the sulfides, selenides and tellurides, and of other semiconductors.

The work is valuable since it represents the thoughts of an investigator who has contributed much through his personal work in the field, and the combination of experimental approach with theoretical analysis is authentic and shows with completeness yet with economy the problems and progress in an important area of physics. Its length is such that there is a minimum of superfluous material. The book appears carefully written, and but a few typographical errors were found. The review of applications of the Hall effect is, unfortunately, confined to one or two pages. The book is too specialized to make a good text, but will have value as a reference work.

W. C. DUNLAP

J. ROBERTS: **High Frequency Applications of Ferrites.** Van Nostrand, New York, 1960. 166 pp., $4.85.

ALTHOUGH the title of this book implies that the author is concerned with the use of ferrites at high frequencies, he touches on applications which range from d.c. magnets to high-speed switching cores. To write a book which treats in detail the physics of ferrites and their many applications for such a frequency range would require several thousand pages. In such a treatment the basic physical principles would probably be obscured in mathematical detail and descriptions of specific device geometries. Professor ROBERTS has written a book which avoids the possible shortcomings of a long book. His book is primarily intended for the newcomer to the ferrite field and requires a background comparable to that of an advanced student of electrical engineering and physics. He uses mathematical analysis sparingly and then only as a tool to illustrate the basic physical principles. For each type of application, whether it be microwave or inductor circuitry, the author proceeds from the pertinent physical principles to several illustrations of specific devices and a discussion of important design factors. Throughout the book he makes an effort to make a comparison between appropriate materials, bringing out the assets of each material and showing the steps approaching improved device performance.

The opening chapter deals with the physical properties of ferrites. This includes the composition and crystallographic structure of the normal and inverted spinels, their resistivity, mechanical and thermal properties, and their magnetic structure. In his discussion of the source of ferrite magnetism, Professor ROBERTS delves into exchange and superexchange interactions and into Neel's theory of ferrimagnetism. The excursion is very short and yet is sufficient to indicate the nature of ferrimagnetism as it is understood today. There are also descriptions of magnetic anisotropy and the gyromagnetic phenomena.

Inductor and transformer cores are the subject of the second chapter. Here the author spends some time and analysis illustrating the value of the air gap in improving the Q of a core and the temperature coefficient of effective permeability. He then discusses initial permeability of a number of materials and follows with a description of several devices such as the pot core inductor and the shell-type transformer core. An effort is made to show not only how performance depends on the initial permeability and loss tangent of the ferrite material, but also upon the enclosing electrical circuit. In the examples, all these factors are considered and the result is a compromise which depends on the particular application.

The following chapter discusses the usefulness of the gyromagnetic properties of ferrites. After short derivations of the complex microwave susceptibility, the ferromagnetic-resonance phenomena, plane-wave propagation in an infinite medium, and Faraday rotation, the author proceeds to describe microwave devices such as isolators, circulators, phase shifters, harmonic generators and ferromagnetic amplifiers. The presentation is by no means exhaustive and serves to outline the types of functions ferrites can perform in the microwave region. Reference is made to the use of these components in larger systems and the need for specific material properties is frequently mentioned. In this chapter the reviewer became particularly aware of one of the possible shortcomings of a survey book of this nature. That is, a reader may be left with the impression that some of the devices such as a waveguide phase shifter

are very easy to construct. It would be well to mention how empirical much of the design work to date has been and how the construction of a device with high performance characteristics is frequently more an art than a science. The reviewer, however, feels that this is really only a minor criticism.

The last two chapters deal with computer applications and with hexagonal materials of high anisotropy. As a result of a square hysteresis loop, high-speed switching cores are used in computers as storage elements. The important parameters such as coercive force, squareness, remnant flux density and switching time are described and two examples of core arrays, the shifting register and the matrix storage, are given. The concluding chapter describes applications of high-anisotropy materials in permanent magnets, pulse transformers and high-frequency microwave devices.

In summary, the reviewer feels that the chief value of this book is its usefulness as an introduction to ferrites and their applications. The book is a survey and as such is short, clear and concise. Nevertheless there is sufficient presentation of physical principles and engineering approach that should satisfy both the physicist and the electrical engineer. There are many devices that have been described and a rich literature exists from which more specialized information can be obtained.

JEROME J. GREEN

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 79–90. Printed in Great Britain

A NEW METHOD OF ZONE LEVELING FOR MATERIALS HAVING DISTRIBUTION COEFFICIENTS IN THE RANGE OF 0–0·6

I. BRAUN and E. Y. WANG

Raytheon Research Division, Waltham, Massachusetts

(*Received* 14 *March* 1961; *in revised form* 13 *April* 1961)

Abstract—By doping an originally pure crystal at both ends and by subsequently melting it repeatedly back and forth, the calculated impurity distributions, as obtained after each pass, converge to a horizontal straight line. This is, therefore, a useful method for obtaining a uniform impurity distribution along a single crystal. In the last zone, however, the abrupt change in impurity concentration, due to the normal freezing cannot be avoided. The first doping of the crystal has to be done *before* the first pass and the second doping at the other end of the crystal *after* the first pass. Graphs are given for different values of k, for the length of the crystal in relation to the zone length and for the minimum number of passes needed to obtain a uniform impurity distribution. The results indicate that the method is extremely well suited for materials with values of k up to 0·6, while ordinary zone leveling fails for materials with a value of k between 0·1 and 1. In the new method the impurity content in any pass is proportional to the introduced impurity content at one end. Thus any desired uniform impurity distribution can be easily obtained.

Résumé—En dopant un cristal originellement pur aux deux extrêmités et en le fondant ensuite continuellement en mouvement d'avance et de recul, les valeurs des distributions d'impuretés calculées obtenues après chaque passage convergent vers une ligne droite horizontale. C'est donc une méthode utile pour obtenir une distribution uniforme d'impuretés le long d'un seul cristal. Dans la dernière zone on ne peut toutefois empêcher le changement abrupt de la concentration d'impuretés dû à la congélation normale. La première dope du cristal doit être faite *avant* le premier passage et la seconde dope à l'autre extrêmité *après* le premier passage. Des graphiques sont donnés pour différentes valeurs de k, pour la longueur du cristal relative à la longueur de la zone, et pour le nombre de passages minimum nécessaires pour obtenir une distribution uniforme d'impuretés. Les résultats indiquent que la méthode est très indiquée pour des matériaux ayant des valeurs de k jusque 0,6, tandis que le nivellement ordinaire de zone ne réussit pas pour des matériaux ayant des valeurs k entre 0,1 et 1. Dans la nouvelle méthode, le contenu d'impuretés de tout passage est proportionnel au contenu d'impuretés introduit dans une extrêmité. Donc, on peut obtenir n'importe quelle distribution d'impuretés désirée.

Zusammenfassung—Wird ein ursprünglich reiner Kristall an beiden Enden dotiert und dann wiederholten Durchgängen durch Zonenschmelze unterworfen, so nähern sich die Störstellenverteilungen, wenn sie nach jedem Durchgang berechnet werden, einer horizontalen Geraden. Man erhält also auf diese Weise eine gleichförmige Störstellenverteilung längs eines Einkristalls. In der letzten Zone bleibt jedoch die durch das normale Erstarren verursachte plötzliche Veränderung der Störstellenkonzentration bestehen. Die erste Dotierung des Kristalls muss *vor* dem ersten Durchgang erfolgen, die zweite am anderen Ende *nach* dem ersten Durchgang. Es folgen graphische Darstellungen für verschiedene Werte von k, für die Länge des Kristalls in Verhältnis zur Zonenlänge und für die minimale Anzahl von Durchgängen, die zur Erzielung der gleichförmigen Störstellenverteilung erforderlich sind. Die Methode ist offenbar für Materialien mit k-Werten bis zu 0,6 sehr geeignet, während sich bei Materialien mit k-Werten zwischen 0,1 und 1 kein Zonenausgleich ergibt. Bei der neuen Methode ist der Gehalt an Verunreininigungen bei jedem Durchgang zu dem an einem Ende eingeführten Betrag an Verunreinung proportional, so dass sich jede gewünschte gleichförmige Störstellenverteilung leicht erreichen lässt.

I. INTRODUCTION

THE purpose of zone leveling is to obtain a uniform impurity distribution. For small values of k (about 0·1) the method of doping at one end and then passing the molten zone along the crystal is effective,[1] but more passes are needed than with the method to be described. PFANN's suggestion to reduce the zone length continuously during melting, so that the impurity concentration remains constant, is difficult to realize experimentally. ROMANEKO[2] describes a method of doping at one end and melting back and forth. This technique can be used with increasing difficulty up to $k = 0{\cdot}25$ and the leveling is far from perfect.

The suggested method is experimentally simple and effective up to $k = 0{\cdot}6$, as will be shown. An originally pure crystal has to be doped at both ends with the same amount of impurities and must then be melted repeatedly back and forth a certain number of times. We shall designate by C_t the solid impurity concentration at each end of the crystal. It is experimentally easier to dope the impurities into the liquid zone rather than to attach them in solid form to both ends. Thus the doping has to be performed in two stages. *Before* melting is initiated the liquid impurity concentration C_t/k has to be introduced into the starting zone. The second doping with the same amount C_t/k has to be performed at the other end *after* the first melting pass. This is because doping into the liquid zone performed at *both* ends before the first pass will lead to diffusion of the impurities of the last zone n into part of the $n-1$ zone during melting. Thus a certain asymmetry is introduced and this complicates the mathematical solution of the problem. After the first pass there is no need to go through the normal freezing in the last zone. Once the front of the zone reaches the end of the ingot the direction of zone travel must be reversed.

By moving the molten zone from left to right, x-co-ordinates have to be used, and, by moving backwards, u-co-ordinates have to be used with $u = L-x$, where L is the total length of the crystal (see Fig. 1). The letter C is used to indicate any impurity concentration and the subscript N the number of passes. Thus, $C_N[x]$ is the impurity concentration at the point x of the Nth pass when the zone moves from left to right and $C_{N+1}[u]$ is the impurity concentration at the point $u = L-x$ in the subsequent pass. Fig. 1 represents schematically the impurity concentrations of two successive passes. To avoid confusion square brackets are used to indicate *only* the argument of the function. We indicate by l the length of the molten zone.

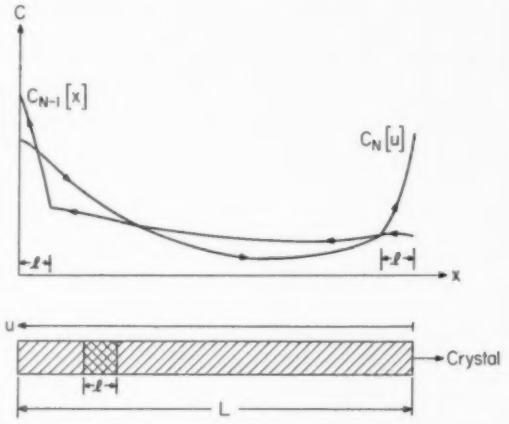


FIG. 1. Schematical representation of two successive impurity concentrations. The arrows indicate the direction of melting.

Using PFANN's[3] assumptions and theory, the impurity concentration of the Nth pass at every point along the crystal is obtained by the integration of the equation

$$\frac{dC_N[x]}{dx} = \frac{k}{l}\{C_{N-1}[x+l]-C_N[x]\} \qquad (1)$$

for $0 \leqslant x \leqslant L-l$, if the zone moves from left to right. In the last zone, normal freezing takes place and the impurity distribution is given by

$$C_N[x] = C_N[x = L-l]\left(\frac{L-x}{l}\right)^{k-1} \qquad (2)$$

for $L-l \leqslant x < L$. If the zone moves from right to left, equations (1) and (2) are used again with the substitution of x by u.

II. CALCULATION OF THE DIFFERENT PASSES

A. $N = 1$

After introducing the impurity concentration C_t at the point $x = 0$ and moving the zone from left to right, and by assuming an originally pure

crystal, i.e. $C_0 = 0$, intregration of equation (1) with the boundary condition $C_1[x = 0] = C_l$ yields

$$C_1[x] = C_l \exp\{-(k/l)x\} \tag{3}$$

for $0 \leqslant x \leqslant L-l$, and for the last zone, equation (2) yields

$$C_1[x] = C_1[x = L-l]\left(\frac{L-x}{l}\right)^{k-1} \tag{4}$$

for $L-l \leqslant x < L$ with

$$C_1[x = L-l] = C_l \exp\{-(k/l)(L-l)\}, \tag{5}$$

from equation (3).

B. $N = 2$

Before melting back in the opposite direction, doping has to be performed again at $x = L$ or $u = 0$ of the amount of C_l/k into the liquid zone. The last zone which had just undergone normal freezing during the first pass will be molten again. This gives the boundary condition for the second pass at the starting point $u = 0$, i.e. the impurity concentration at this point is equal to the sum of the freshly introduced impurity concentration C_l and the remaining impurity due to the normal freezing from the first pass

$$C_2[u = 0] = C_l + \frac{k}{l}\int_0^l C_1[u]\,du \tag{6}$$

In equation (6), C_1 $[u]$ has to be taken for the normal freezing curve given by equations (4) with (5), and by changing the co-ordinates with $x = L-u$, equation (6) yields, after integration,

$$C_2[u = 0] = C_l + C_l \exp\{(-k/l)(L-l)\} \tag{7}$$

To find the impurity distribution along the crystal, equation (1) has to be integrated for $N = 2$ and replacing x by u, which yields with the boundary condition (7)

$$C_2[u] = C_l \exp\{-(k/l)u\} + \frac{C_l}{2} \times \left\{\exp\{(-k/l)(L-l+u)\} + \exp\{(-k/l)(L-l-u)\}\right\} \tag{8}$$

for $0 \leqslant u \leqslant L-l$. To calculate equation (8), the expression $C_1[u+l]$ was needed. This expression has been obtained by changing the co-ordinates with $x = L-u$ in equation (3), which yields $C_1[u] = C_l \exp\{(-k/l)(L-u)\}$ and then by replacing $[u]$ by $[u+l]$, i.e.

$$C_1[u+l] = C_l \exp\{(-k/l)(L-l-u)\} \tag{9}$$

is obtained.

C. $N = 3, 4, \ldots N$

After the second pass, melting has to be performed backwards and forwards without introducing further impurities. If N is odd, the zone moves from left to right; if N is even the zone moves in the opposite direction. As boundary condition for any $N > 2$, the following equation is used

$$C_N[x = 0] = C_{N-1}[x = l] = C_{N-1}[u = L-l] \tag{10}$$

if the molten zone moves from left to right. Equation (10) holds again if the zone moves from right to left, only x has to be replaced by u and vice versa. The proof of equation (10) is as follows. At the starting point $x = 0$ for the Nth pass, towards the right, the impurity concentration $C_N[x = 0]$ is due to the normal freezing of the $(N-1)$th pass in the opposite direction, i.e.

$$C_N[x = 0] = \frac{k}{l}\int_0^l C_{N-1}[x]\,dx \tag{11}$$

with

$$C_{N-1}[x] = C_{N-1}[x = l]\left(\frac{x}{l}\right)^{k-1} \tag{12}$$

If equation (12) is introduced into equation (11) and this equation integrated, we get the boundary conditions (10).

To obtain $C_3[x]$, equation (1) has to be integrated for $N = 3$ which yields

$$C_3[x] = k/l \exp\{-(k/l)x\} \int \times C_2[x+l] \exp\{(k/l)x\}\,dx + K \exp\{-(k/l)x\} \tag{13}$$

for $0 \leqslant x \leqslant L-l$. $C_2[x+l]$ must be calculated from equation (8) by first changing the variables, with $u = L-x$, and then by changing the argument $[x]$ into $[x+l]$. The value of K must be determined with the boundary condition (10). Introducing the value of K so obtained and $C_2[x+l]$ into equation

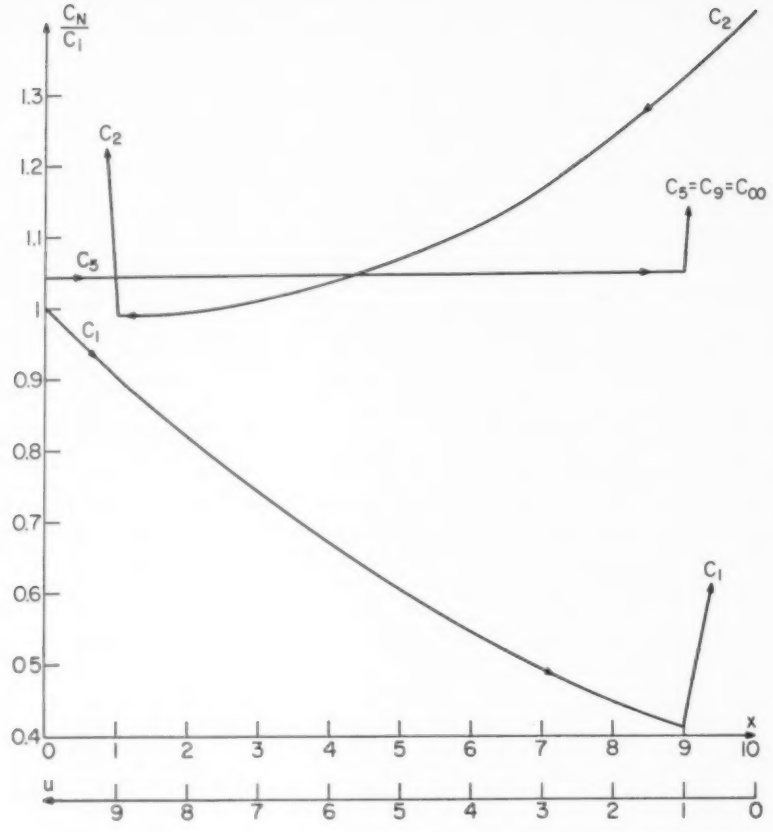


FIG. 2. $l = 1$ cm, $L = 10$ cm, $k = 0{\cdot}1$. At least five passes are needed to obtain a uniform impurity distribution. $C_9[x]$ is very close to $C_5[x]$, $C_\infty = 1{\cdot}05\ C_t$.

(13), the impurity distribution of the third pass can be written in the form

$$\frac{C_3[x]}{C_t} =$$

$$= \sum_{\mu=0}^{2} p_3^{\mu}[\mu(L-l)+x] \exp\{(-k/l)(\mu(L-l)+x)\}$$

$$+ \sum_{\mu=1}^{2} p_3^{\mu}[\mu(L-l)-x] \exp\{(-k/l)(\mu(L-l)-x)\} \tag{14}$$

for $0 \leqslant x \leqslant L-l$, where we introduced as abbreviation

$$\left.\begin{aligned} P_3^0[x] &= \tfrac{1}{2}+\tfrac{1}{2}\frac{k}{l}x \\ P_3^1[L-l-x] &= P_3^1[L-l+x] = P_3^1 = \tfrac{1}{2} \\ P_3^2[2(L-l)-x] &= P_3^2[2(L-l)+x] = P_3^2 = \tfrac{1}{4} \end{aligned}\right\} \tag{15}$$

which will facilitate a generalization for any N. If the P_3^i-functions reduce to a constant, it is meaningless to write the argument of the function as for instance $P_3^2 = \frac{1}{4}$ in equation (15).

Finally the Nth impurity concentration can be calculated with the knowledge of the $(N-1)$th impurity concentration by integrating equation (1) for $N = N$ and by using the boundary conditions of equation (10). The result of a rather long

calculation is

$$\frac{C_N[u]}{C_i} = \sum_{\mu=0}^{N-1} P_N^{\mu}[\mu(L-l)+u] \exp\{(-k/l)[\mu(L-l)+u]\} + \sum_{\mu=1}^{N-1} P_N^{\mu}[\mu(L-l)-u] \exp\{(-k/l)[\mu(L-l)-u]\} \tag{16}$$

The same expression holds if the zone moves in the opposition direction, except that u has to be replaced by x. $P_N^i[i(L-l)+u]$ and $P_N^i[i(L-l)-u]$ stand for finite power series of their arguments and they are of the form

$$P_N^i[i(L-l)+u] = A_{N,0}^i + A_{N,1}^i\{i(L-l)+u\} + A_{N,2}^i\{i(L-l)+u\}^2+\ldots$$
$$P_N^i[i(L-l)-u] = A_{N,0}^i + A_{N,1}^i\{i(L-l)-u\} + A_{N,2}^i\{i(L-l)-u\}^2+\ldots \tag{17}$$

It can be shown that the coefficients $A_{N,j}^i$ of both series of equation (17) have the same value. These series are designated together by $P_N^i[i(L-l) \pm u]$,

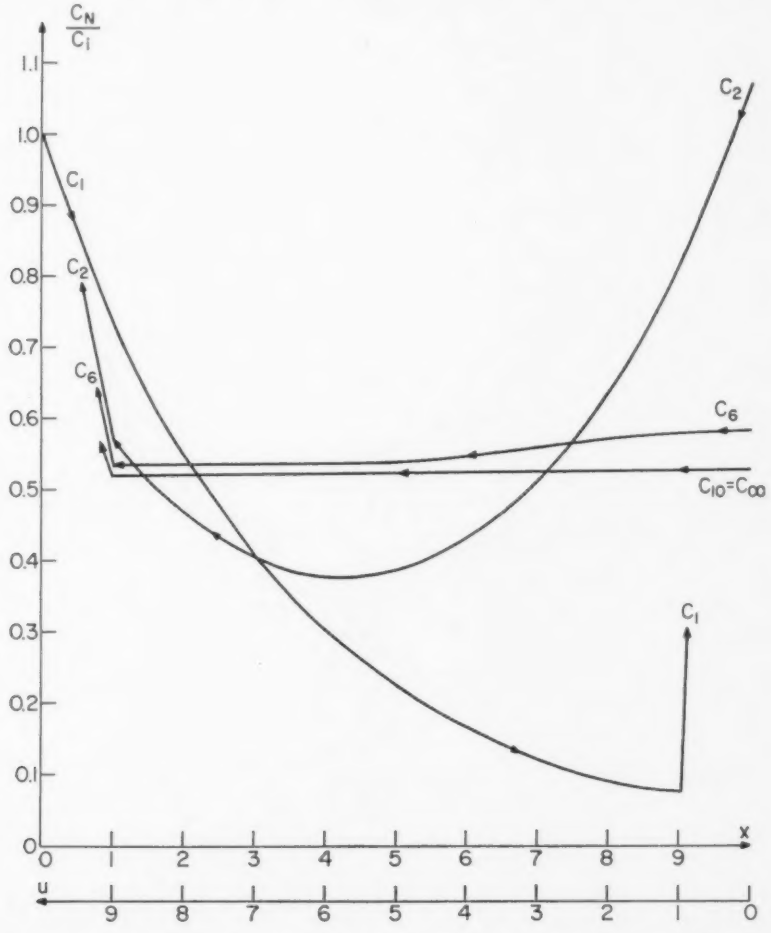


FIG. 3. $l = 1$ cm, $L = 10$ cm, $k = 0{\cdot}3$. At least ten passes are needed to obtain the uniform impurity distribution. $C_\infty = 0{\cdot}52\ C_i$.

where the upper index i is always a summation index and the lower index N indicates the number of passes. The P^i_N-functions of the Nth pass must be calculated with the help of the P^i_{N-1}-functions of the $(N-1)$th pass by the following formulas (the proof is straightforward but long and has not been given here).

(1) *For* $i > 0$

$$P^i_N[i(L-l) \pm u] = \tfrac{1}{2}\{P^{i-1}_{N-1}[i(L-l) \pm u] + \frac{l}{2k} P^{(1)i-1}_{N-1}[i(L-l) \pm u] + \left(\frac{l}{2k}\right)^2 P^{(2)i-1}_{N-1}[i(L-l) \pm u] + \ldots\} \tag{18}$$

where $P^{(1)i-1}_{N-1}[i(L-l) \pm u]$, $P^{(2)i-1}_{N-1}[i(L-l) \pm u]$, ... are taken to mean the first, second ... derivative of $P^{i-1}_{N-1}[i(L-l) \pm u]$ in respect to its argument $[i(L-l) \pm u]$.

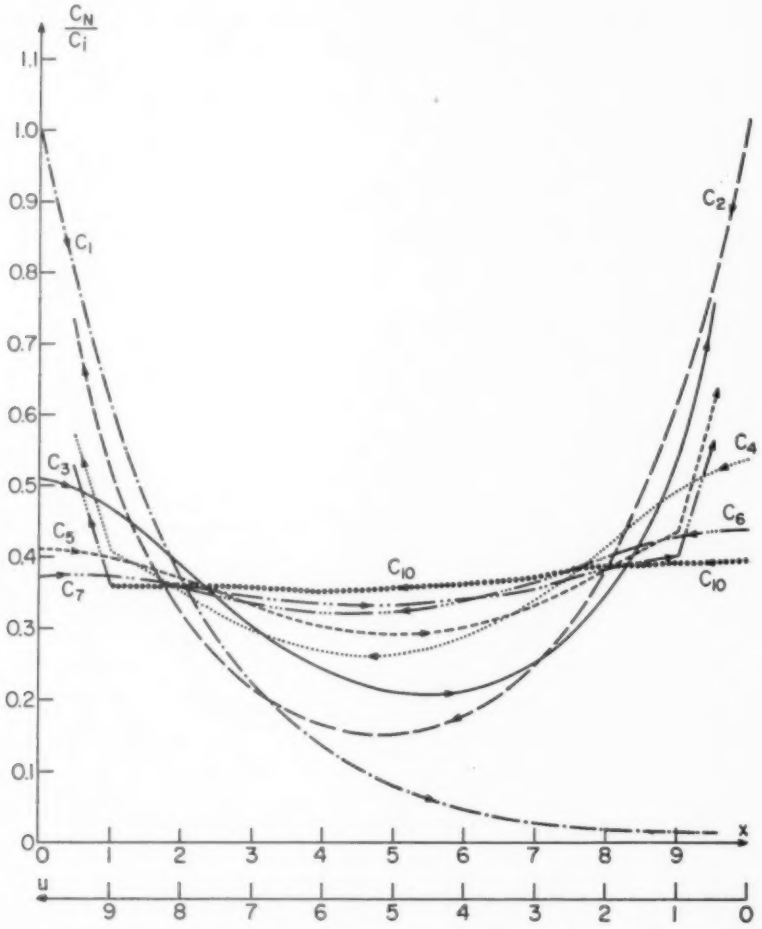


FIG. 4. $l = 1$ cm, $k = 0{\cdot}5$. Graphical representation of the impurity distributions of ten successive passes, which converge slowly to an ultimate impurity distribution. Ten passes are not enough to reach the ultimate concentration distribution. It is suggested that at least fourteen passes are needed. $C_\infty = 0{\cdot}36$ $C_1 = C_{14}$.

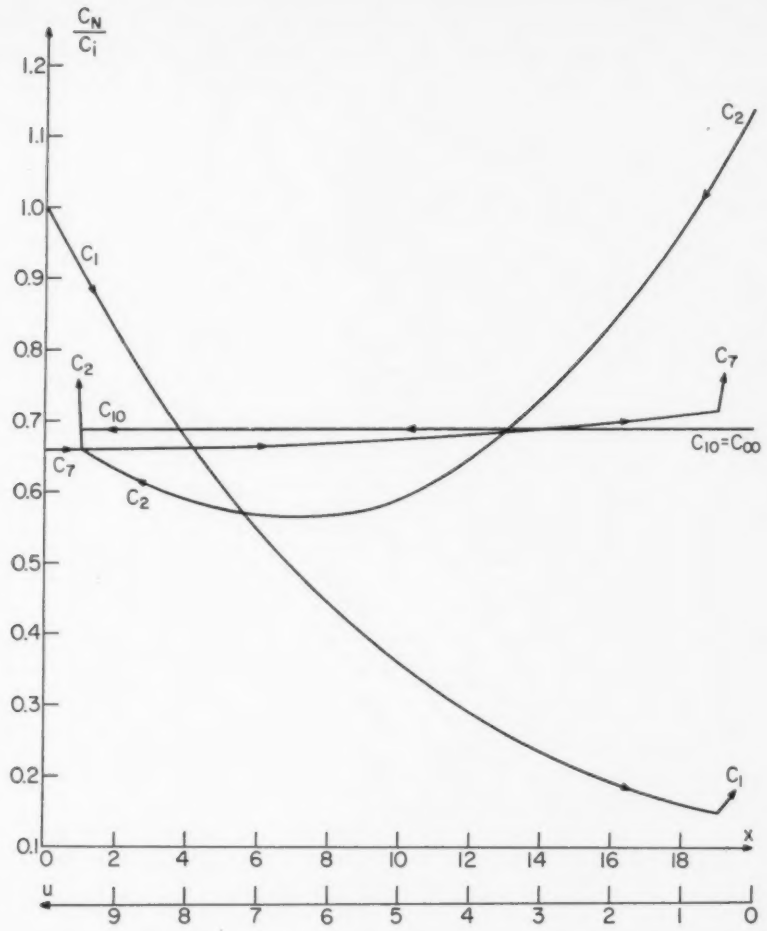


FIG. 5. $l = 1$ cm, $L = 20$ cm, $k = 0{\cdot}1$. Ten passes are needed. $C_\infty = 0{\cdot}39$ $C_i = C_{10}$.

(2) *For* $i = 0$

$$P_N^0[u] = \frac{k}{l}\int P_{N-1}^1[u]\,\mathrm{du} + P_{N-1}^1[u=0] \tag{19}$$

Note, that in accordance with equation (16) $P_N^0[-u]$ does not exist. If the zone moves from left to right, the same expressions (18) and (19) have to be used with u substituted by x.

A detailed example will be given in the Appendix, together with the different P_N^i-series up to $N = 7$.

III. GRAPHICAL REPRESENTATION OF THE DIFFERENT IMPURITY CONCENTRATIONS

In accordance with equation (16) it is seen that every impurity concentration C_N is proportional to the introduced impurity concentration C_i at one end of the crystal. This proportionality makes it possible to increase or decrease the uniform impurity concentration along the crystal in an experimentally convenient way. The different curves have been plotted independently of C_i, i.e. in the form of C_N/C_i.

Different groups of graphs were plotted, holding

the zone length l constant and varying k and the length of the crystal L. Figs. 2–4 show the case of $l = 1$ cm, $L = 10$ cm and k varying from 0·1 to 0·5. It can be seen from these figures that, with more passes are needed by increasing the whole length of the crystal L.

A small number of passes N is needed if k and L in relation to l are small, for example, for $k = 0{\cdot}1$,

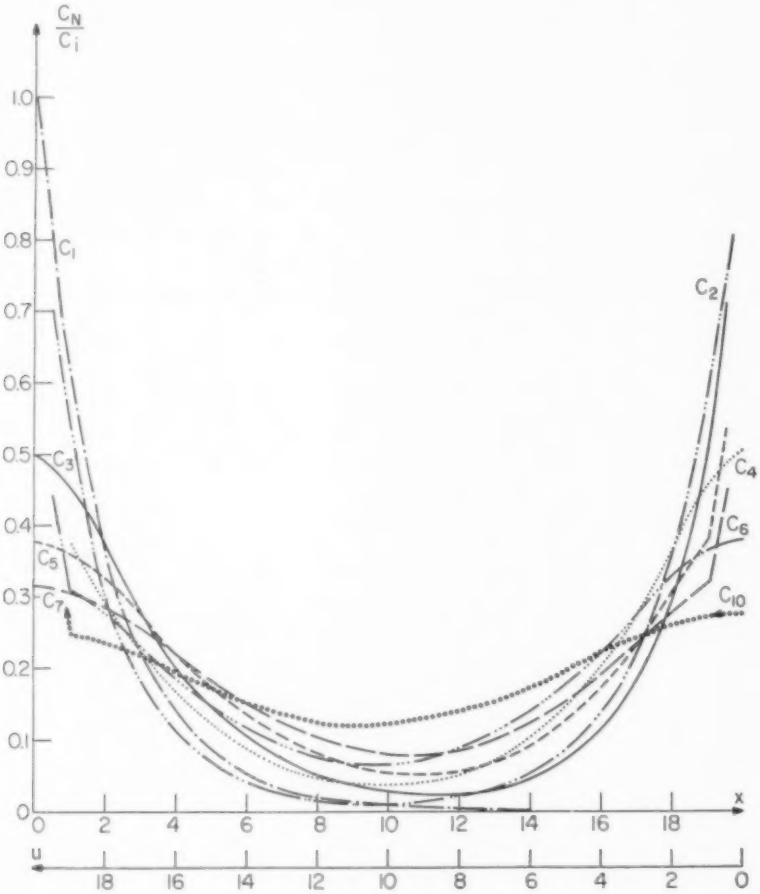


FIG. 6. $l = 1$ cm, $L = 20$ cm, $k = 0{\cdot}5$. Ten passes are not enough to reach the ultimate impurity distribution. It is not certain, but probable, that this ultimate distribution is uniform along the crystal. $C_\infty = 0{\cdot}19\, C_i$.

increasing k, and L and l remaining constant, an increasing number of passes are needed to reach the impurity concentration which approximates a horozontal line, which had been designated by C_∞. The same can be seen for the case of k varying again from 0·1 to 0·5 and $l = 1$ cm, $L = 20$ cm as in Fig. 5. Holding k and l constant and varying L we see also by comparing Fig. 2 with Fig. 5 that $l = 1$ cm, $L = 10$ cm, only five passes are needed, but if k increases to 0·5 we need up to fourteen passes.

Figs. 4, 6 and 7 demonstrate clearly that if the number of passes increases, a limiting—or ultimate—impurity distribution is reached in the sense that if N increases $C_N[u] = C_{N+2}[u]$ and $C_{N-1}[x] = C_{N+1}[x]$; this ultimate impurity distri-

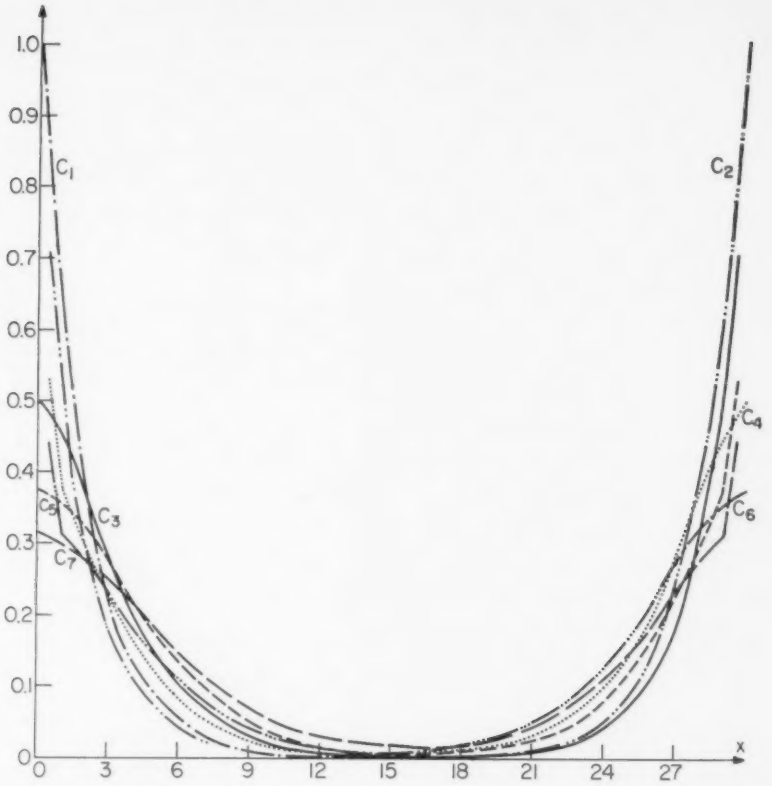


FIG. 7. $l = 1$ cm, $L = 30$ cm, $k = 0{\cdot}5$. Seven successive impurity distributions are plotted. It is doubtful whether the ultimate impurity distribution will approach a straight line.

bution is in certain ranges of l, L and K a straight horizontal line, but certainly not for any value of l, L and k. See, therefore, Fig. 7 with $l = 1$ cm, $L = 30$ cm and $k = 0{\cdot}5$ where it is doubtful that the ultimate distribution will flatten out.

IV. CONCLUSION

It is concluded that the zone-leveling method described leads to a uniform impurity distribution along the whole crystal—except for the last zone—for a certain range of L, in relation to l and k. The equalization of the impurity concentration is reached with a smaller number of passes if k and L are small for $l = 1$ cm. Since it is experimentally easier to adjust the length of the crystal L to a given zone length l than to adjust l to L for a given distribution coefficient k, Table 1 gives the maximum length L_{max} the crystal may have in relation to $l = 1$ cm for certain values of k to assure a uniform impurity concentration.

Table 1*

L_{max} (cm)	k
8	0·6
10	0·5
15	0·3
20	0·2
40	0·1

* $l = 1$ cm.

If the number of melting passes increases, the

curves of the impurity concentrations converge for increasing N to a limiting or ultimate impurity distribution which is only a straight line for certain values of L, l and k. The ultimate impurity distribution is in reality oscillating, since every melting process will bring the impurities from the first zone—due to the normal freezing of the previous pass—into the end zone. In the region of $l \leqslant x \leqslant L-l$ or $l \leqslant u \leqslant L-l$, however, the ultimate impurity distribution will remain the same. Mathematically this ultimate impurity distribution can be defined by $C_{N-1}[x] = C_{N+1}[x]$ or $C_N[u] = C_{N+2}[u]$.

The method described is probably not the only possible way to obtain a nearly uniform impurity distribution. Spot doping at constant intervals may eventually give a uniform impurity distribution. However, the mathematical treatment yields a non-analytical function for the impurity concentration which changes its form from interval to interval for the same pass. Thus for any given set of values for k, L and l it is difficult to predict the number of passes needed to equalize the impurity concentration.

Acknowledgements.—The authors are grateful for many helpful discussions with Dr. J. PATEL and for the help received from Miss B. KILEY with the numerical calculations.

REFERENCES

1. W. G. PFANN, *Zone Melting* p. 153. Wiley, New York (1958).
2. V. N. ROMANEKO, *Fiz. Tvergodo* Tela, **1**, 11, 1679 (1959).
3. W. G. PFANN, *Trans. Amer. Inst. Min. (Metall.) Engrs.* **194**, 747 (1952).

APPENDIX

1. *Detailed example*

As an example, $C_5\,[x]$ will be calculated with the help of equation (16) knowing that $C_4[u]$ is of the form

$$\frac{C_4[u]}{C_i} =$$

$$= \sum_{\mu=0}^{3} P_4^{\mu}[\mu(L-l)+u] \exp\left[(-k/l)\{\mu(L-l)+u\}\right]$$

$$+ \sum_{\mu=1}^{3} P_4^{\mu}[\mu(L-l)-u] \exp\left[(-k/l\{\mu(L-l)-u\}\right] \tag{A.1}$$

for $0 \leqslant u \leqslant L-l$, with

$$P_4^0[u] = \tfrac{1}{2}+\tfrac{1}{2}\frac{k}{l}u \tag{A.2}$$

$$P_4^1[L-l \pm u] = \tfrac{3}{8}+\frac{k}{4l}(L-l \pm u) \tag{A.3}$$

$$P_4^2[2(L-l) \pm u] \equiv P_4^2 = \tfrac{1}{4} \tag{A.4}$$

$$P_4^3[3(L-l) \pm u] \equiv P_4^3 = \tfrac{1}{8} \tag{A.5}$$

The different P_4^i-series must be calculated with the help of equations (18) and (19) knowing the relations P_4^i. Since the 5th pass is from left to right (N is an odd number), x-co-ordinates have to be used and equation (19) will give for $N = 5$

$$P_5^0[x] = \frac{k}{l}\int P_4^1[x]\,dx + P_4^1[x=0] \tag{A.6}$$

$P_4^1[x]$ can be calculated with the help of (A.3), simply by replacing $[L-l \pm u]$ by $[x]$

$$P_4^1[x] = \tfrac{3}{8}+\frac{k}{4l}x$$

and for $x = 0$

$$P_4^1[x=0] = \tfrac{3}{8}$$

Introducing these values into (A.6):

$$P_5^0[x] = \tfrac{3}{8}+\tfrac{3}{8}\frac{k}{l}x+\frac{k^2}{8l^2}x^2 \tag{A.7}$$

P_5^1 must be calculated with equation (18) for $N = 5$, $i = 1$:

$$P_5^1[L-l \pm x] = \tfrac{1}{2}\{P_4^0[L-l \pm x] + \frac{1}{2k}P_4^{(1)0}[L-l \pm x]\} \tag{A.8}$$

To calculate for this expression $P_4^0[L-l \pm x]$ the argument $[u]$ of (A.2) has to be exchanged by $[L-l \pm x]$ which yields

$$P_4^0[L-l \pm x] = \tfrac{1}{2}+\tfrac{1}{2}\frac{k}{l}[L-l \pm x]$$

and by taking the derivatives:

$$P_4^{(1)0}[L-l \pm x] = \tfrac{1}{2}\frac{k}{l}$$

$$P_4^{(2)0}[L-l \pm x] = 0$$

and by introducing these values into (A.8) yields:

$$P_5^1[L-l \pm x] = \tfrac{3}{8}+\frac{k}{4l}(L-l \pm x)$$

Table A.1. Different P_j^i-functions up to $N = 7$

N	P_N^0	P_N^1	P_N^2	P_N^3	P_N^4	P_N^5	P_N^6	P_N^7
1	1	0	0	0	0	0	0	0
2	1	$\frac{1}{2}$	0	0	0	0	0	0
3	$\frac{1}{2}+\frac{1}{2}\frac{k}{l}x$	$\frac{1}{2}$	$\frac{1}{4}$	0	0	0	0	0
4	$\frac{1}{2}+\frac{1}{2}\frac{k}{l}u$	$\frac{3}{8}+\frac{k}{4l}(L-l\pm u)$	$\frac{1}{4}$	$\frac{1}{8}$	0	0	0	0
5	$\frac{3}{8}+\frac{3}{8}\frac{k}{l}x+\frac{k^2}{8l^2}x^2$	$\frac{3}{8}+\frac{k}{4l}(L-l\pm x)$	$\frac{1}{4}+\frac{k}{8l}(2L-2l\pm x)$	$\frac{1}{8}$	$\frac{1}{16}$	0	0	0
6	$\frac{3}{8}+\frac{3}{8}\frac{k}{l}u+\frac{k^2}{8l^2}u^2$	$\frac{5}{16}+\frac{k}{4l}(L-l\pm u)+\frac{k^2}{16l^2}(L-l\pm u)^2$	$\frac{1}{4}+\frac{k}{8l}(2L-2l\pm u)$	$\frac{5}{32}+\frac{k}{16l}(3L-3l\pm u)$	$\frac{1}{16}$	$\frac{1}{32}$	0	0
7	$\frac{5}{16}+\frac{5}{16}\frac{k}{l}x+\frac{k^2}{8l^2}x^2+\frac{k^3}{48l^3}x^3$	$\frac{5}{16}+\frac{k}{4l}(L-l\pm x)+\frac{k^2}{16l^2}(L-l\pm x)^2$	$\frac{15}{64}+\frac{5}{32}\frac{k}{l}(2L-2l\pm x)+\frac{k^2}{32l^2}(2L-2l\pm x)^2$	$\frac{5}{32}+\frac{k}{16l}(3L-3l\pm x)$	$\frac{3}{32}+\frac{k}{32l}(4L-4l\pm x)$	$\frac{1}{32}$	$\frac{1}{64}$	0

In a similar manner again from equation (18) for N = 5, $i = 2$

$$P_5^2[2(L-l) \pm x] = \tfrac{1}{2}\{P_4^1[2(L-l) \pm x] + \frac{l}{2k} P_4^{(1)1}[2(L-l) \pm x]\} \tag{A.9}$$

$P_4^1[2(L-l) \pm x]$ is obtained from (A.3) by exchanging $[L-l \pm u]$ by $[2(L-l) \pm x]$, and by taking the derivatives yields

$$\left.\begin{aligned} P_4^1[2(L-l) \pm x] &= \tfrac{1}{4} + \frac{k}{4l}\{2(L-l) \pm x\} \\ P_4^{(1)1}[2(L-l) \pm x] &= \frac{k}{4l} \\ P_4^{(2)1}[2(L-l) \pm x] &= 0 \end{aligned}\right\}$$

Introducing these values P_5^1 is obtained with

$$P_5^2[2(L-l) \pm x] = \tfrac{1}{4} + \frac{k}{8l}(2L-2l \pm x) \tag{A.10}$$

In a similar manner, application of equation (18) with $N = 5$ and $i = 3, 4$ yields

$$P_5^3 = \tfrac{1}{8}$$

$$P_5^4 = \tfrac{1}{16}$$

Knowing the different P_5^i-functions, $C_5[x]$ is obtained with equation (16)

$$\frac{C_5[x]}{C_i} =$$

$$= (\tfrac{3}{8} + \tfrac{3}{8}\frac{k}{l}x + \frac{k^2}{8l^2}x^2) \exp\{(-k/l)x\}$$

$$+ \left(\tfrac{3}{8} + \frac{k}{4l}(L-l+x)\right) \exp\{(-k/l)(L-l+x)\}$$

$$+ \left(\tfrac{3}{8} + \frac{k}{4l}(L-l-x)\right) \exp\{(-k/l)(L-l-x)\}$$

$$+ \left(\tfrac{1}{4} + \frac{k}{8l}(2L-2l+x)\right) \exp\{(-k/l)(2L-2l+x)\}$$

$$+ \left(\tfrac{1}{4} + \frac{k}{8l}(2L-2l-x)\right) \exp\{(-k/l)(2L-2l-x)\}$$

$$+ \tfrac{1}{8}\Big\{\exp\{(-k/l)(3L-3l+x)\} + \exp\{(-k/l)(4L-4l-x)\}\Big\}$$

$$+ \tfrac{1}{16}\Big\{\exp\{(-k/l)(4L-4l+x)\} + \exp\{(-k/l)(4L-4l-x)\}\Big\}$$

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 91–99. Printed in Great Britain

THE DESIGN THEORY OF THERMOELECTRIC COOLING ELEMENTS AND UNITS

J. E. PARROTT and A. W. PENN

Research Laboratory, Associated Electrical Industries, Aldermaston, Berkshire

(*Received* 10 *March* 1961)

Abstract—In order to assist in the design of thermoelectric refrigerators, the relationship between refrigerating capacity per unit volume and coefficient of performance is examined for a single-stage unit. These are both determined for given temperature difference by the figure of merit and the contact resistance, the latter being particularly important in fixing the maximum value of refrigerating capacity per unit volume. Equations and graphs are presented, which, when certain considerations of an economic character have been assessed, will enable an optimum design to be arrived at.

Résumé—Pour pouvoir faciliter la dimensionnement des appareils frigorifiques thermoélectriques, la relation entre la capacité réfrigérante par unité de volume et le coefficient de rendement est examinée pour un dispositif à un étage. On les détermine pour une différence de température donnée par le facteur de mérite et la résistance de contact, cette dernière étant particulièrement importante pour fixer la valeur maximum de la capacité réfrigérante par unité de volume. Des équations et des graphiques sont presentés pour permettre d'obtenir un dimensionnement optimum quand certaines considérations d'ordre économique doivent être estimées.

Zusammenfassung—Zur Unterstützung der Fabrikation thermoelektrischer Kühler wurde die Beziehung zwischen Kühlungskapazität pro Volumeinheit und Wirkungsgrad an einem einstufigen Gerät untersucht. Bei gegebener Temperaturdifferenz werden beide durch die Leistungsziffer und den Übergangswiderstand bestimmt, wobei der letztere zur Festsetzung des Höchstwertes der Kühlungskapazität pro Volumeinheit besondere Bedeutung hat. Die angegebenen Gleichungen und graphischen Darstellungen sollten, nach Bestimmung gewisser wirtschaftlicher Gesichtspunkte, die Herstellung eines optimalen Gerätes ermöglichen.

1. INTRODUCTION

THE physics of thermoelements has received considerable attention over the past fifty years, and is now fairly well understood. The geometry of these elements, excluding the problem of heat transfer from the junctions, has, however, received little, if any, study. This disparity is unreasonable, since one may gain considerably in efficiency and refrigerating capacity by correct design, as well as by obtaining materials which are potentially capable of greater cooling.

The intention of this paper is to provide the formulae, etc., needed by the designer in carrying out his task. To this end we shall calculate (for various values of the figure of merit of the material, the contact resistances and the temperature difference) the geometry and the operating conditions of the thermoelements for which the refrigerating capacity per unit volume and/or coefficient of performance are optimized.

2. NOTATION

ω	dimensionless figure of merit
Π	Peltier coefficient
κ	(total) thermal conductivity
ρ	resistivity
T	absolute temperature—and temperature of hot junctions
S	Seebeck coefficient
r_c	contact resistance/unit area
q	thermal load per cm^2
ΔT	temperature difference across the thermocouple
j	current density

A	cross-sectional area
l	length of thermocouple arms
1, 2 (subs.)	refer to arm 1 and arm 2
f	$\Delta T/2T$
β	$r_c/\rho l$
$\varkappa$	$j\rho l/2\Pi(1+\beta)$
ψ	$q/2l$ refrigerating capacity per unit volume
ψ_{max}	maximum refrigerating capacity per unit volume with respect to current
ψ_c	maximum refrigerating capacity per unit volume for maximum efficiency mode of operation
ϵ_{ψ}	efficiency (total) for the case of maximum refrigerating capacity per unit volume
ϵ_{max}	maximum efficiency with respect to current
j_m	j for the maximum temperature difference
C	cost of unit refrigeration
Γ	$[\omega/4(1+\beta)-f]\beta^2$
h	heat leakage coefficient

3. MODEL AND BACKGROUND

The usefulness of a material for the purposes of thermoelectric cooling may be defined in terms of a figure of merit

$$\omega = \frac{\pi^2}{k\rho T} \tag{1}$$

Various other definitions have been adopted,[1-3] but these differ only in constants and the power of the temperature; thus, inter-conversion for the purposes of comparison is a simple matter.

The dependence of the figure of merit upon various physical parameters of the materials has been discussed in detail by JOFFE and others,[1, 4] and recent papers have reported various materials with figures of merit up to approximately 1.[5, 6]

The application of this principle for the purpose of small-scale refrigeration is increasing, and some devices, such as the continuously recording dew-point hygrometer, are already available.[7-9] MARTYNOVSKY and NAER[10] have discussed some of the problems of heat transfer associated with such devices.

A refrigerating thermocouple unit is shown schematically in Fig. 1. The arms are composed of two bars of semiconductor of Seebeck coefficient S and $-S$, resistivity ρ and thermal conductivity κ. We shall suppose their length to be l and that a contact resistance r_c per unit area exists between the arms at the cold junction where a thermal load of

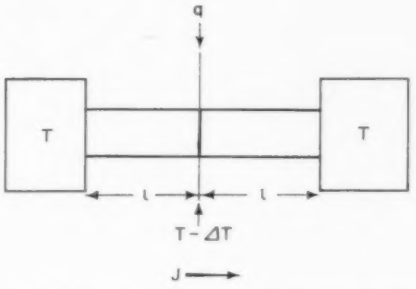


FIG. 1. Schematic diagram of a thermoelectric refrigerator.

q W/cm^2 is applied. A temperature difference of ΔT is supposed to exist when the hot junctions are maintained at a temperature T and a current density of j is flowing in the appropriate direction.

For simplicity we have assumed equal thermal and electrical resistivities, cross-sectional areas, and lengths. Where the thermal and electrical resistivities are not equal, the length and cross-sectional areas must be adjusted so that

$$\frac{\rho_1 l_1^2}{\kappa_1 A_1^2} = \frac{\rho_2 l_2^2}{\kappa_2 A_2^2} \tag{2}$$

in order to obtain the best performance.

The heat balance at the cold junction, using the simple model, is given by the following equation

$$q = 2\Pi j - (\rho l + r_c) j^2 - \frac{2\kappa}{l}\Delta T \tag{3}$$

where on the left-hand side we have the refrigerating load, and on the right-hand side we have first the Peltier heat absorption and then, subtracted from that, half of the resistive joule heating in the conductor (the other half is effectively absorbed at the outer contacts), plus the amount of heat conducted along the elements from the hot junctions. The refrigerating capacity is therefore greatest for any given current when ΔT is zero (that is, we are using the device purely as a heat

pump), and is zero when

$$\Delta T = \frac{2\Pi j - (\rho l + r_c) j^2}{2\kappa/l}$$

This gives what is essentially a load curve, shown in Fig. 2; q is a heat leak which is usually proportional to ΔT, say $h\Delta T$, and the intersection of the two curves gives the refrigerating capacity and temperature difference for a given current.

The dependence of q on j means that q will first increase with current, go through a maximum, and then decrease with further increase of current. The current at which this maximum effect is produced is

$$j_m = \frac{\Pi}{\rho l + r_c}$$

and at this current

$$q = \frac{\Pi^2}{\rho l + r_c} - \frac{2\kappa \Delta T}{l}$$

The refrigerating capacity is zero when

$$\Delta T = \Pi^2 / 2\kappa\rho(1 + r_c/\rho l)$$

which is therefore the maximum reduction in temperature which can be produced.

The power consumed (W) in the thermocouple is the resistance heating plus work done in driving the current through the thermocouple against the Seebeck voltage caused by ΔT. Thus, defining the coefficient of performance as refrigerating capacity divided by power consumed, we find

$$\epsilon = \frac{q}{W} = \frac{2\Pi j - (\rho l + r_c) j^2 - 2\kappa \Delta T/l}{2j[(\rho l + r_c) j + \Pi \Delta T/T]} \qquad (4)$$

Equations (3) and (4) form the basis on which the theory of the succeeding sections is erected.

4. STATEMENT OF PROBLEM, FOR A SINGLE-STAGE THERMOELECTRIC REFRIGERATOR

The designer of any useful device is restricted by two sets of factors; on the one hand the specification for the device, and on the other the available material and technology. In the case of a thermoelectric refrigerator, the first class includes:

(1) The temperature difference.
(2) The refrigerating load.

The principal factors in the second class are:

(a) The figure of merit of the materials available.
(b) The contact resistance between the thermoelements for such materials.

Within the limitations set out above, the designer will be guided by two considerations of a basically economic character:

(1) To reduce the capital cost of the refrigerator, he will desire to minimize the quantity of semiconductor needed.

(2). To reduce the running cost, he will wish to make the refrigerator as efficient as possible. Put more precisely, he will wish to obtain the best refrigerating capacity per unit volume ψ and highest coefficient of performance ϵ, bearing in mind that these are not independent. The relative importance of these two parameters depends on detailed economic factors.*

The designer must first consider the relative merits of single- and multi-stage (cascade) refrigeration. We shall not discuss this question at all,

---

* The object of the designer is generally to meet the specification (temperature difference and refrigerating capacity) as cheaply as possible; i.e. the cost per unit of refrigeration should be as low as possible. This is

$$C = \frac{\text{Capital cost}}{\text{Average refrigerating capacity}} \times \left[\frac{1}{\text{Life}} + \text{Interest rate}\right] + \frac{\text{Cost of power}}{\text{Coefficient of performance}}$$

This expression contains many factors which we cannot take into account, e.g. the capital cost will include, in addition to the semiconductor and its fabrication, the cabinet power supply, etc. Also the over-all coefficient of performance will usually be lower than the figure we have calculated. However, if we can separate these terms in the cost which depend on the variables at our disposal, β and x, from the rest, the problem will be simplified. We suggest that this can be done approximately with an equation of the form

$$C = \frac{A}{\psi(\beta, X)} + \frac{B}{\epsilon(\beta, X)} + D$$

where A, B and D depend only weakly on β and x. A corresponds in general terms to the cost per unit volume of the thermoelectric unit and B to the cost of electric power.

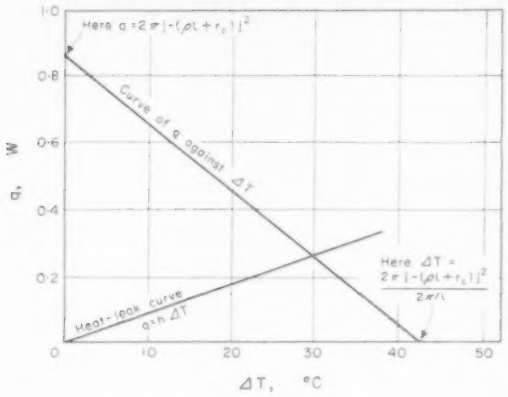


FIG. 2. Load curve for a thermoelectric refrigerator.

as it involves very great complications. Assuming, therefore, that only a single-stage refrigerating unit is being considered, the two variables at the disposal of the designer are the lengths of the thermocouple arms and the operating current. He wishes to adjust these to optimize his design with respect to the considerations given in the previous paragraph; when he has done this, it will be seen from (3) that the load curve of q against ΔT is now defined completely. The specified temperature difference then determines the refrigerating capacity per unit area, and the specified load fixes the total cross-section required. This means that the fundamental design problem has been solved.

5. DESIGN THEORY FOR A SINGLE-STAGE THERMOELECTRIC REFRIGERATOR

The refrigerating capacity per unit volume $\psi = q/2l$, and, using (3), is given by

$$\psi = \frac{1}{2l}\left[2\Pi j - (\rho l + r_c)j^2 - \frac{2\kappa\Delta T}{l}\right] \qquad (5)$$

Making the substitutions

$$f = \frac{\Delta T}{2T}, \beta = \frac{r_c}{\rho l}, x = \frac{j\rho l}{2\Pi}(1+\beta)$$

gives us

$$\psi = \frac{2\kappa\rho^2 T}{r}\beta^2\left[\frac{\omega}{1+\beta}x(1+x) - f\right] \qquad (6)$$

This equation shows that given two materials of the same ω and β, that having the larger value of $\kappa\rho^2$ is preferable. The designer is given f; he can vary β and x independently through l and j.

The design can be optimized for refrigerating capacity per unit volume in two ways. This is most easily seen by regarding the current (or x) as the variable and the element length (or β) as a parameter. For any given length of element (value of β) there is a certain current density (value of x) which corresponds to the greatest refrigerating capacity for that particular geometry of element. The corresponding value of ψ is naturally dependent on the geometry and will be a maximum for a particular value of β; this is the maximum possible refrigerating capacity per unit volume obtainable from material of given figure of merit and with given contact resistance. This is the first approach.

In the second approach we consider first the coefficient of performance. Again for any given length of element there is a particular current

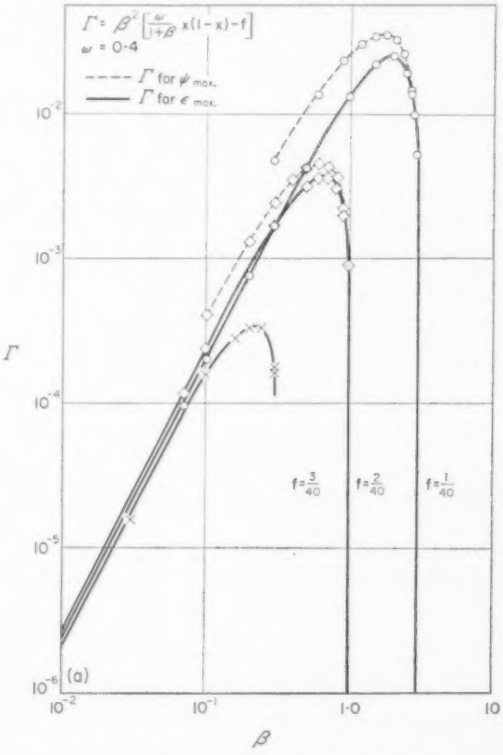


FIG. 3(a). Refrigerating capacity per unit volume.

density which corresponds to the highest coefficient of performance for that element; this current density is always less than in the first case and so the refrigerating capacity per unit volume is less. Nevertheless there is again a particular value of β which corresponds to the maximum value of this resulting refrigerating capacity.

Because the coefficient of performance increases monotonically with increase of length of element (decrease of β), a third approach in which its absolute maximum is sought does not really exist.

We can now apply the ideas outlined above to the expression for the refrigerating capacity.

Examination of (6) shows that ψ is a maximum when $x = \frac{1}{2}$ and is given by

$$\psi_{\max} = \frac{2\kappa\rho^2 T}{r_c^2}\beta^2\left[\frac{\omega}{4(1+\beta)} - f\right] \tag{7}$$

$$= \frac{2\kappa\rho^2 T}{r_c^2}\Gamma(\omega,\beta,f) \tag{8}$$

defining Γ. The variation of Γ with β for certain values of f and ω is shown in Fig. 3 (a–c). The values of f and ω cover most probable operating conditions and figures of merit.

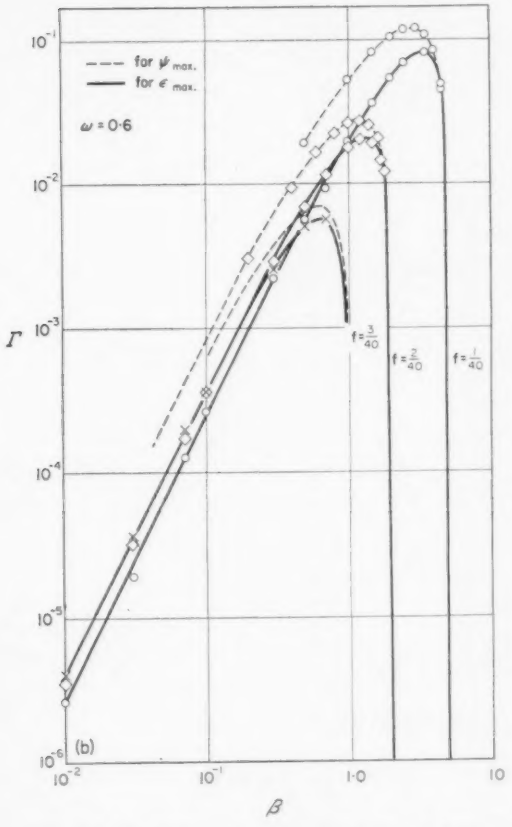


FIG. 3(b). Refrigerating capacity per unit volume.

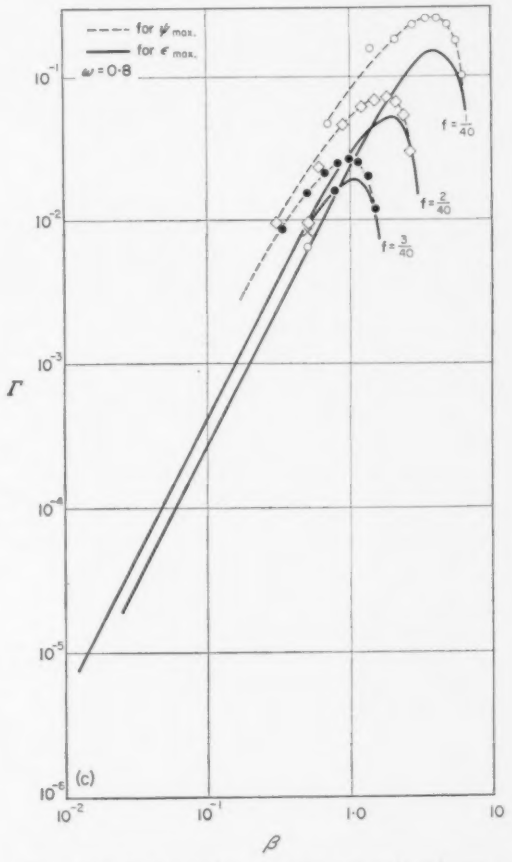


FIG. 3(c). Refrigerating capacity per unit volume.

We now use (4) and calculate the efficiency for this mode of operation. Substituting f, β and x into (4) gives us

$$\epsilon = \frac{x - x^2 - f(1+\beta)/\omega}{2x(x+f)} \tag{9}$$

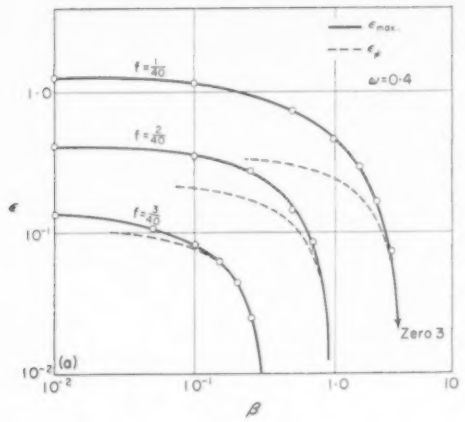


FIG. 4(a). Coefficient of performance.

and setting $x = \frac{1}{2}$ to obtain maximum refrigerating capacity and writing $\epsilon_\psi = \epsilon$ at this point

$$\epsilon_\psi = \frac{\omega - 4f(1+\beta)}{2\omega(1+2f)} \tag{10}$$

Fig. 4 (a–c) shows the variation of ϵ_ψ with β for the same values of f and ω as were used in Fig. 3 (a–c).

The second alternative is to operate the thermocouple so that the coefficient of performance has its maximum value. The current for this condition is obtained by differentiating ϵ with respect to x and setting $\partial\epsilon/\partial x = 0$. The value of x so obtained is

$$x = \frac{f}{[1+\omega(1+f)/(1+\beta)]^{1/2}-1}$$

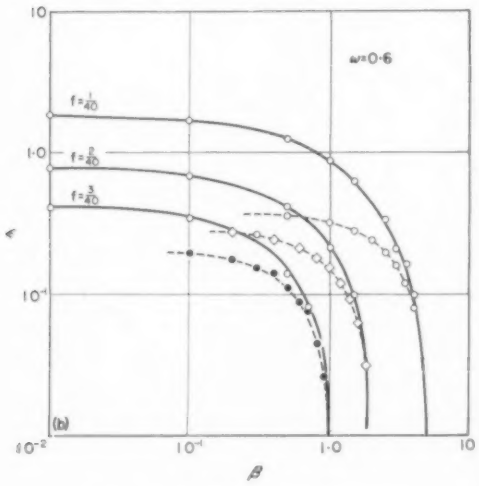


FIG. 4(b). Coefficient of performance.

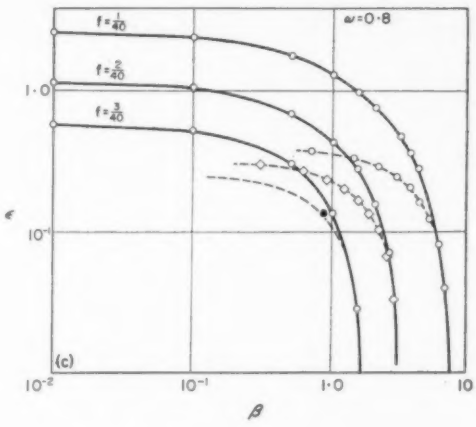


FIG. 4(c). Coefficient of performance.

and in Fig. 5 the dependence of x on β is shown for the previously used values of f and ω.

Inserting this expression for x in equations (6) and (9) gives us the refrigerating capacity per unit volume.

$$\psi_\epsilon = \frac{2\kappa\rho^2 T}{r_e^2}\beta^2 f$$

$$\times\left[\frac{\omega}{1+\beta}\frac{\{[1+\omega(1+f)/(1+\beta)]^{1/2}-1-f\}}{\{[1+\omega(1+f)/(1+\beta)]^{1/2}-1\}^2}-1\right] \tag{11}$$

and the coefficient of performance

$$\epsilon_{max} = \frac{1}{2f}\left[\frac{[1+\omega(1+f)/(1+\beta)]^{1/2}-(1+2f)}{[1+\omega(1+f)/(1+\beta)]^{1/2}+1}\right] \tag{12}$$

The values of Γ and ϵ obtained from (11) and (12) are also plotted in Fig. 3 (a–c) and Fig. 4 (a–c).

It will be noticed that the notation used is that ϵ_ψ means the value of ϵ when ψ is maximized and ψ_ϵ means the value of ψ when ϵ is maximized.

Looking at the figures, it will be seen that whilst ψ has a maximum at a certain value of β, ϵ increases as β decreases. They both go to zero at

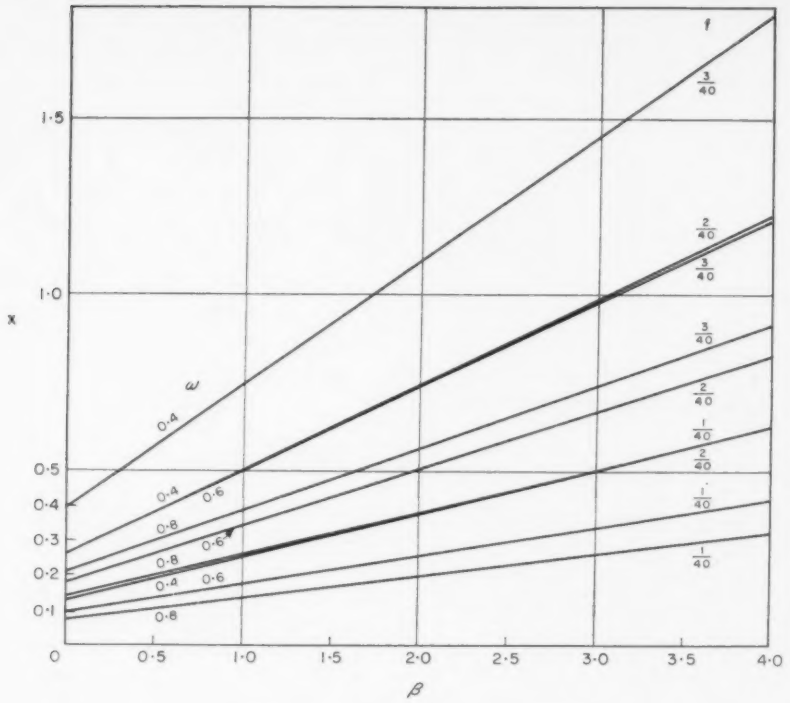


FIG. 5. Optimum refrigerating current.

the same value of β, however, irrespective of whether ψ or ϵ has been maximized. The physical significance of this can be seen from equation (3). As the contact resistance increases (β increases), q will decrease towards zero; however, q cannot become negative in a real situation. What really happens is that the required ΔT will not be obtainable. As long as q is positive for the required ΔT, the refrigerator can be constructed if a large enough cross-section is used. The distinction between maximizing ψ or ϵ disappears under these circumstances near the limit of operation.

Fig. 6 (a–c) has been constructed from Figs. 3 and 4 and shows Γ plotted in terms of ϵ for the two types of maximizing. These show what values of ψ and ϵ can be obtained when ω and f have been fixed. From these curves we see that Γ goes through a maximum, and therefore the designer will always work on the right-hand side of this maximum to get the larger value of coefficient of performance. Also the curves for ψ maximized

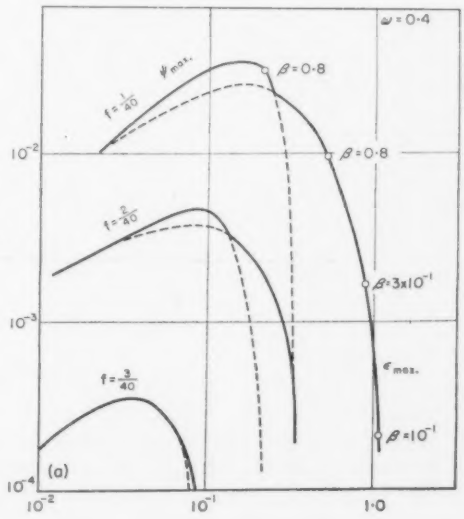


FIG. 6(a). Design curves.

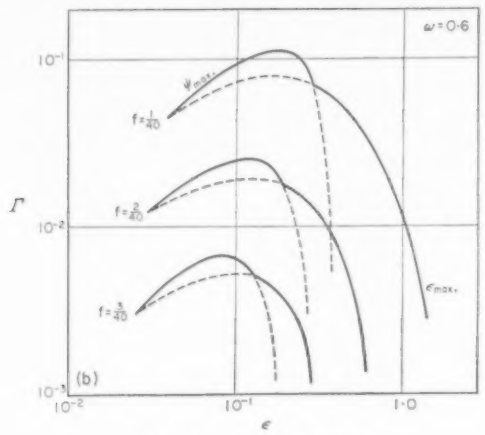


FIG. 6(b). Design curves.

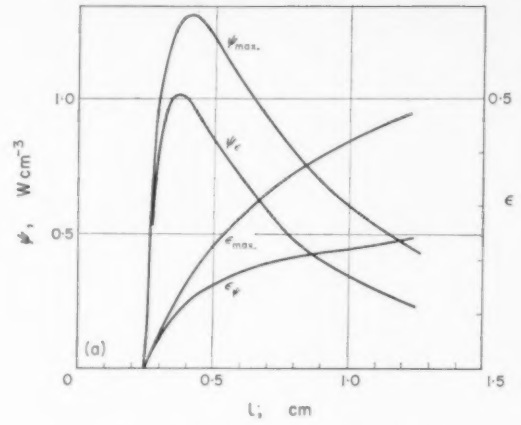


FIG. 7(a). Specimen performance curves.

and ϵ maximized cross over one another. It will always be best to use the higher of the two curves, and in fact for some efficiencies or refrigerating capacities only one is possible. The graphs also

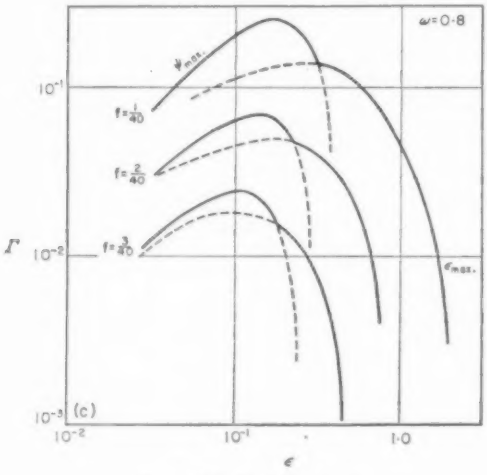


FIG. 6(c). Design curves.

show that for larger values of f the curves move closer together.

The actual values of ψ and ϵ to be chosen can only be arrived at after taking economic considerations (see footnote p. 93) into account. However, it certainly appears that a very much higher β can be tolerated in many cases than had hitherto been expected. In fact, for higher figures of merit and lower temperature difference, it is advantageous to make r_c rather greater than ρl.

A concrete example is shown in Fig. 7(a), where ψ and ϵ are plotted against length of thermocouple arms for reasonable values of the material parameter and contact resistance. Despite the fairly large contact resistance, the maximum refrigerating capacity is found with a length of about 4 mm where r_c is a little larger than ρl. In practice, the designer would certainly choose to operate at a somewhat larger value of l in order to increase the efficiency.

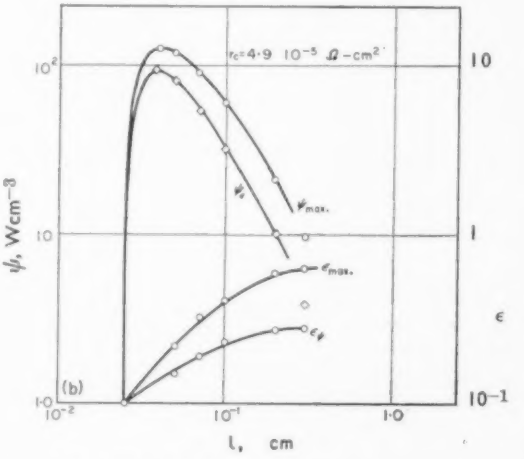


FIG. 7(b). Specimen performance curves.

A reduction in contact resistance makes possible a very great increase in refrigerating capacity. Fig. 7(b) is for the same material as Fig. 7(a) but with r_c reduced by an order of magnitude. There is in consequence an increase of ψ by two orders of magnitude, ϵ remaining, however, almost unchanged. The designer will here undoubtedly sacrifice some refrigerating capacity for increased coefficient of performance, particularly as l for the maximum of ψ is now inconveniently small at about $\frac{1}{2}$ mm.

REFERENCES

1. A. Joffe, *Semiconducting Thermoelements and Thermoelectric Cooling*. Infosearch, London (1957).
2. J. E. Parrott, A.E.I. Research Report No. A.899 (1959).
3. R. P. Chasmar, M–V Report No. 10642 (1955).
4. R. P. Chasmar and R. Stratton, *J. Electron. Control* **7,** 52 (1959).
5. *Ohio Semiconductors—Data Sheet* (1959).
6. H. J. Goldsmid and A. R. Sheard, *Brit. Pat.* 834593 (1960).
7. *Electron. Ind.* **17,** 10, 93 (1958).
8. F. K. Iordanishvili and L. G. Thanlich, Zh. tekh. Fiz., **27,** 1215, (1957).
9. P. Gerthsen, J. A. A. Gilsing and M. Van Tol, *Philips tech. Rev.* **21,** 196 (1960).
10. V. S. Martynovsky and V. A. Naer, *Proceedings of the Tenth International Congress of Refrigeration*, Paper 3–34. Institut International du Froid, Paris (1959).

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 100–104. Printed in Great Britain

HEAT TO A.C. CONVERSION USING THERMOELECTRIC GENERATORS AND TUNNEL DIODES

F. G. GEVERT,* S. WANG* and R. U. BROZE†

(*Received* 3 *March* 1961; *in revised form* 3 *April* 1961)

Abstract—Single-cell lead telluride thermoelectric generators when operated between 600 and 50°C generate approximately 1W of electric power. However, the output voltage of the generator for a matched load is of the order of 120 mV and the current is of the order of 8 A. The output of such a single-cell thermoelectric generator can be converted into high-voltage low-current power by using a high-current tunnel diode as a d.c. to a.c. converter.

Résumé—Les générateurs thermoélectriques des cellules simples au tellurure de plomb produisent environ un watt de puissance électrique dans la gamme de température de 600 à 50°C. Pourtant, la tension de sortie du générateur pour une charge égalée est de l'ordre de 120 mV. et le courant est de l'ordre de 8 A. La tension de sortie d'un tel générateur thermoélectrique peut être convertie en une haute tension et un faible courant en employant comme convertisseur de courant continu à courant alternatif une diode tunnel.

Zusammenfassung—Einzellige thermoelektrische Generatoren aus Bleitellurid erzeugen bei Verwendung zwischen 600 und 50°C etwa 1 W elektrische Energie. Die Ausgangsspannung des Generators beträgt bei geeigneter Belastung 120 mV, und der Strom ist von der Grössenordnung 8 A. Die Leistung eines derartigen einzelligen thermoelektrischen Generators lässt sich in einen hochgespannten Niederstrom umwandeln, wenn man eine hochstromige Tunneldiode als Wechselrichter benutzt.

INTRODUCTION

THIS paper explores the possibility of using tunnel diodes in conjunction with thermoelectric generators to convert the low-voltage output of a single thermoelectric cell into practical higher voltages. The tunnel diode appeared most suitable because the voltage required to bias the tunnel diode into the negative-resistance region[1] is obtainable from single-cell lead telluride thermoelectric generators. Since the thermoelectric cell is inherently a low-impedance power source, a high-current tunnel diode must be used in order to achieve efficient d.c. to a.c. conversion.

At the time the work was started, commercial tunnel diodes were only available with peak currents of the order of 20 mA. This would leave two ways to make high-current tunnel diodes: either to use a very large number of low-current tunnel diodes in parallel or to investigate the technology of producing higher-current tunnel-diode junctions.

The purpose of the work reported here is to show that tunnel diodes used in conjunction with thermoelectric generators can be used successfully as heat to a.c. converters.

THE THERMOELECTRIC GENERATOR[2]

A thermoelectric generator as illustrated in Fig. 1 was built. It consists of a heater—*a*; a ceramic plate—*b*; a high-purity iron busbar—*c*; *p*- and *n*-type lead telluride elements ($\frac{1}{4}$-in. diameter by $\frac{1}{4}$-in. length)—*d*; two copper busbars—*e*; a water-cooled heat sink—*f*; and four screws—*g*. The contact between the lead telluride and the iron busbar is made by pressing the two materials together in a special jig at 600°C in an argon atmosphere. This contact cannot be soldered because it is the high-temperature contact, which is to be maintained at 600°C. Silver solder and other high-

* University of California, Berkeley, California.

† Fairchild Semiconductors, Inc., San Rafael, California.

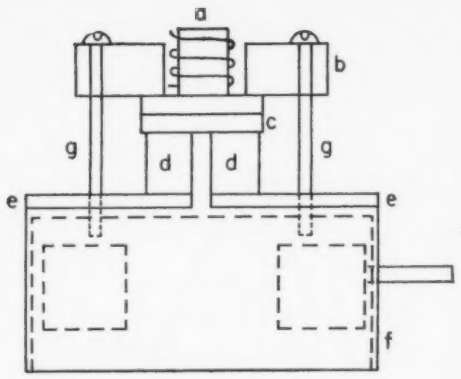


FIG. 1. Single-cell lead telluride thermoelectric generator.

temperature solders contain copper, a metal that diffuses rapidly into lead telluride, changing its properties. The contact between the lead telluride and the copper is a soldered contact. Lead telluride can be tinned with lead–tin solder by means of an ultrasonic soldering iron.

The thermoelectric device must be operated in an inert atmosphere in order to minimize the oxidation of the lead telluride elements and the iron busbar. The characteristics of the generator when operated between 600 and 50°C are the following:

Open-circuit voltage:	240 mV
Internal resistance:	13 mΩ

TUNNEL-DIODE FABRICATION[3]

Because of the prohibitive cost and the possibility of variation in bias voltage (due to differences in base spreading resistance), the use of a large number of low-peak-current tunnel diodes in parallel was abandoned in favor of producing junctions of much higher current. Junctions were fabricated by alloying boron-doped aluminum ($N_A \simeq 1\times10^{21}$ cm^{-3}) to single-crystal phosphorus-doped Si ($N_D = 8{\cdot}2\times10^{20}$ cm^{-3}). The junctions were heated for 2·0 sec in a N_2–H_2 forming gas atmosphere at a temperature of 1050°C.

With these alloying techniques, junctions were produced with a peak current of 360 mA and a maximum peak-to-valley ratio of 3·6. Attempts to alloy larger areas of aluminum–boron alloy than used in forming the 360-mA junction resulted in partial alloying and small shorts which would degenerate the junction.

By alloying several junctions on the same silicon substrate, ultrasonically soldering one large common contact to the back of the substrate, and ultrasonically soldering contacts to the aluminum–boron alloy dots, a parallel-junction tunnel diode having a sufficiently high peak current was formed. The final characteristics obtained were:

Peak-point current:	660 mA
Valley-point current:	230 mA
Peak-point voltage:	65 mV
Valley-point voltage:	340 mV

THE NEGATIVE-RESISTANCE OSCILLATOR

Fig. 2 shows the actual circuit used in this experiment. The voltage of the thermoelectric generator must be high enough to bias the tunnel diode into the negative-resistance region. For most tunnel diodes, the bias must be between 65 and 350 mV. The tank circuit is formed by the inductance of the transformer and the capacitance C, reflected into the primary as n^2C; n is the turns

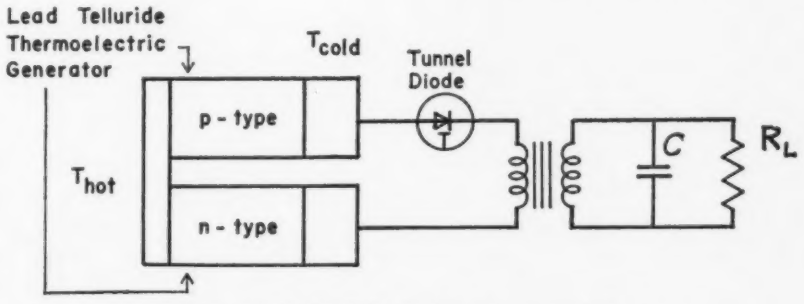


FIG. 2. Actual circuit for heat to a.c. conversion.

ratio of the transformer. The resistance R_{Int} is the total d.c. resistance of the primary circuit. This includes the resistance of the battery, the resistance of the transformer and the resistance of the leads.

Oscillations will occur under the following conditions: (a) The d.c. resistance R_{Int} is smaller than the negative resistance of the tunnel diode. (b) The a.c. resistance is greater than the negative resistance of the tunnel diode. An oscillogram of the output voltage of the oscillator is shown in Fig. 3.

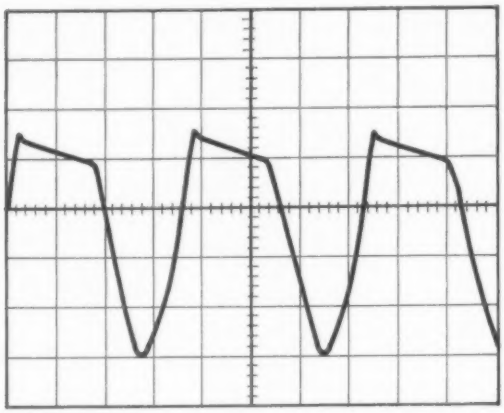


FIG. 3. Waveshape of output voltage (experimental).

THEORETICAL ANALYSIS

For the purpose of analysis, the electrical characteristics of a tunnel diode in and near the negative-resistance region may be approximated by

$$i = -ae+be^3 \tag{1}$$

where i and e are normalized with respect to the quiescent point, that is $i = I-I_0$ and $e = V-V_0$. To obtain a symmetrical a.c. waveform from a negative-resistance oscillator using a tunnel diode, we assume that $V_0 = (V_1+V_2)/2$ and $I_0 = (I_1+I_2)/2$, I_1 and V_1 being the peak-point current and voltage and I_2 and V_2 the valley-point current and voltage, respectively, as illustrated in Fig. 4. By setting $di/de = 0$ and $i = -(I_1-I_2)/2$ at $e = (V_2-V_1)/2$, we obtain

$$a = \tfrac{3}{2}\frac{I_1-I_2}{V_2-V_1} \text{ and } b = 2\frac{I_1-I_2}{(V_2-V_1)^3} \tag{2}$$

The a.c. equivalent circuit of the negative-

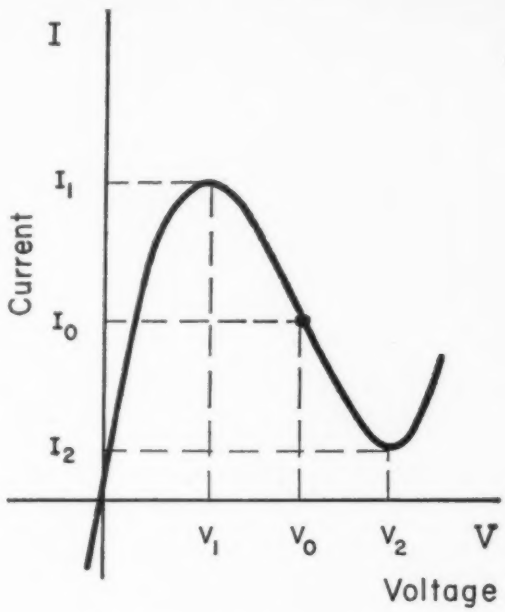


FIG. 4. Tunnel-diode characteristic.

resistance oscillator is shown in Fig. 5. From this circuit we may write

$$i+(i_L+i_C+i_G) = 0$$

or

$$-ae+be^3+\frac{1}{L}\int e\,dt+C\frac{de}{dt}+eG = 0$$

Differentiating and rearranging terms in the above equation, we obtain the standard van der Pol

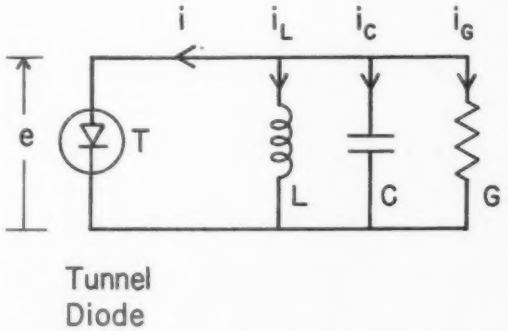


FIG. 5. Equivalent circuit of negative-resistance oscillator.

equation, namely

$$\frac{d^2e}{dt^2}+\alpha(1-\beta)\omega_0\frac{de}{dt}+\omega_0^2e=0 \qquad (3)$$

where $\omega_0^2=1/LC$, $\alpha=(a-G)/\omega_0C$ and $\beta=3b/(a-G)$. For the particular diode used in this experiment, $a=2{\cdot}34$ and $b=42$; and for the present circuit, $\omega_0=2000$, $\alpha=2{\cdot}3$ and $\beta=90$.

The solution of the van der Pol equation is discussed extensively in the literature.[4] For $\alpha\geqslant 1$, the waveshape of e shows an abrupt transition from one state to another and it is often referred to as relaxation oscillation. The voltage e varies with time in the following manner:

$$t=\frac{\alpha}{\omega_0}\left(\ln\frac{\beta^{1/2}e}{2}-\frac{\beta e^2}{2}+2\right) \qquad (4)$$

For $\alpha\leqslant 1$, the solution can be found by perturbation calculation and is given by

$$e=\frac{2}{\beta^{1/2}}\left[\cos\omega_0t+\frac{\alpha}{8}(3\sin\omega_0t-\sin 3\omega_0t)\right] \qquad (5)$$

The waveshape is a flattened and skewed sinusoid.

The waveshapes for $\alpha\gg 1$ and $\alpha\ll 1$ are illustrated in Fig. 6 for comparison with the experimental results. In Fig. 6 two quantities, E_{max} and T, can be checked experimentally. These values can also be obtained from equations (4) and (5). For $\alpha\gg 1$, $E_{max}=2/\beta^{1/2}$ and $T=1{\cdot}61\alpha/\omega_0$; for $\alpha\ll 1$, $E_{max}=2/\beta^{1/2}$ and $T=2\pi/\omega_0$. This gives: $E_{max}=0{\cdot}21$ V (primary side) or $E_{max}=9{\cdot}2$ V (load side); $T=1{\cdot}8\times10^{-3}$ sec for relaxation oscillation and $T=3{\cdot}1\times10^{-3}$ sec for harmonic oscillation. The values observed experimentally are $E_{max}=8{\cdot}5$ V (load side) and $T=3{\cdot}5\times10^{-3}$ sec. The theoretical and experimental values are in good agreement. In the present circuit, the turns ratio of the transformer was 44, and the circuit began to oscillate for $R_L>4000\,\Omega$. (Refer to Fig. 2.) The output voltage changed from 6 to 8·5 V as R_L varied from 4000 to 10,000 Ω.

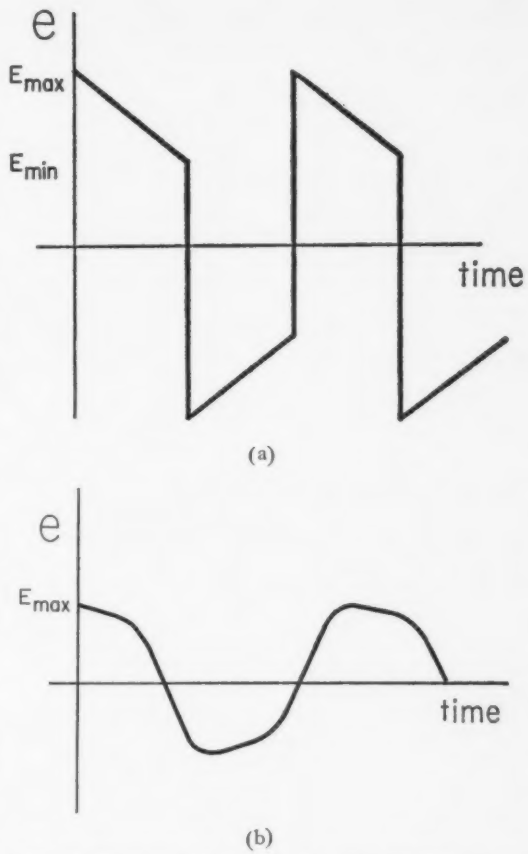


FIG. 6. Waveshape of oscillator output voltage for (a) $\alpha\gg 1$ and (b) $\alpha\ll 1$.

DISCUSSION

It should be pointed out that the i–e relationship given by equation (1) is symmetrical with respect to the quiescent point and hence the predicted waveshape illustrated in Fig. 6 is symmetrical with respect to the horizontal axis. If the diode characteristics are not quite symmetrical with respect to the quiescent point, terms like e^2 and e^4 should be added to equation (1). This explains why the experimental curve shown in Fig. 4 has unsymmetrical positive and negative half-cycles.

In the present investigation, we obtained at the output approximately 9 mW of a.c. power. As mentioned earlier, the thermoelectric generator is a high-current and low-voltage source. To achieve maximum power transfer, a tunnel diode having a current capacity of several amperes is needed. The present state of low output power and

efficiency is due mainly to our inability to make high-current tunnel diodes. A brief discussion concerning the output power and conversion efficiency of a tunnel diode follows.

In the relaxation mode of oscillation, the voltage decreases almost linearly from $E_{max} = 2/\beta^{1/2}$ to $E_{min} = 1/\beta^{1/2}$ as shown in Fig. 6(a). Assuming a linear decrease from E_{max} to E_{min}, we have $E_{rms} = (7/3)^{1/2} E_{min}$ or the a.c. power output

$$P_{a.c.} = E_{rms}^2 G = \frac{7}{9}\frac{(a-G)}{b} G \qquad (6)$$

For maximum $P_{a.c.}$, $G = \frac{1}{2} a$. Thus

$$(P_{a.c.})_{max} = \frac{7}{36}\frac{a^2}{b} = \frac{7}{32}(I_1-I_2)(V_2-V_1) \qquad (7)$$

The d.c. power input into the circuit is

$$P_{d.c.} = I_0 V_0 = \tfrac{1}{4}(I_1+I_2)(V_1+V_2) \qquad (8)$$

and the efficiency of conversion is

$$\eta = \frac{7}{8}\frac{(I_1-I_2)(V_2-V_1)}{(I_1+I_2)(V_2+V_1)} \qquad (9)$$

In equation (7), $V_2 - V_1$ is of the order of 200–400 mV. It is quite easy to have a single cell of thermoelectric generator to supply a current of a few amperes. Thus, a thermoelectric generator in conjunction with a tunnel diode should be able to supply an output power in the watt range if high-current tunnel diodes become available. The conversion efficiency of a tunnel diode depends upon the ratio of the peak-point voltage and current to their respective values at the valley point as given by equation (9). For commercial tunnel diodes, I_1/I_2 can be as high as 40 : 1 and V_1/V_2 is in the range of 3 : 1 to 5 : 1. Thus, a conversion efficiency in the neighborhood of 50 per cent may be achieved.

There are many advantages which tunnel diodes have as a d.c. to a.c. converter, especially if used in conjunction with thermoelectric generators. The *i–e* characteristic of a tunnel diode is not very sensitive to temperature changes. Its operating voltage in the negative resistance region falls in the range of voltage produced by a single thermoelectric cell maintained at a temperature difference of a few hundred degrees. The circuitry involved is quite simple.

REFERENCES

1. L. Esaki, *Phys. Rev.* **109**, 603 (1958).
2. A. F. Joffe, *Semiconductor Thermoelements and Thermoelectric Cooling*. Chaps. I and II. Infosearch, London (1957).
3. I. A. Lesk, N. Holonyak and U. S. Davidsohn, *IRE Wescon Convention Record* pp. 9–31, August (1959).
4. See for example: W. J. Cunningham, *Introduction to Nonlinear Analysis*. Chap. 6. McGraw-Hill, New York (1958).

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 105–109. Printed in Great Britain

ETCH PITS IN GERMANIUM AND THEIR RELATION TO HARDNESS

J. R. DALE and J. C. BRICE

Mullard Research Laboratories, Cross Oak Lane, Salfords, Surrey

(*Received* 22 *September* 1960; *in revised form* 21 *February* 1961)

Abstract—Work is described which shows that the shallow etch pits revealed on germanium {111} faces by etching in CP4 correspond to globular defects which are caused by contamination of the germanium in the liquid state probably by oxygen.

The effect of these defects and of grown-in edge dislocations on hardness has been investigated by means of a microhardness tester. This shows that both defects produce an increase in hardness. The range of hardness values obtained was from 500 to 700 VPN.

Résumé—On décrit les travaux qui démontrent que les creux de corrosion peu profonds qui se produisent sur les faces du germanium {111} en gravant dans le CP4 correspondent à certains défauts globulaires causés par la contamination du germanium dans l'état liquide par l'oxygène.

L'effet de ces défauts ainsi que les dislocations d'arêtes développées sur la dureté a été examiné avec l'aide d'un instrument de mesure de microdureté. Cela démontre que les deux défauts produisent une augmentation de dureté. La gamme des valeurs de dureté obtenues est de 500 à 700 VPN.

Zusammenfassung—Untersuchungen zeitgen, dass die flachen Ätzgruben, die sich auf {111} Germaniumkristallflächen beim Ätzen mit CP4 ergeben, auf kugelförmigen Defekten beruhen. Diese entstehen wahrscheinlich durch Verunreinigung des Germaniums in der flüssigen Phase mit Sauerstoff.

Der Effekt dieser Defekte und wachstumsbedingter Kantenversetzungen auf die Härte wurde mit einem Mikro-Härtestempel untersucht. Es ergibt sich, dass beide Defekte die Härte vergrössern. Die Härte schwankte zwischen den Werten 500 bis 700 der Vickersschen Härteskala.

1. INTRODUCTION

IN THE course of metallurgical work on germanium, the occasional presence of shallow etch pits as well as the normal edge-dislocation-type pits has been noted.[1] TWEET[2] found that large concentrations of small shallow pits occurred in the parts of his crystals last to solidify, while in the part first to solidify there were smaller concentrations of much larger pits. He showed that annealing caused clustering of the small pits to form the larger ones and assumed that these pits were due to vacancy clusters. In this laboratory it had been noted that the density of shallow pits appeared to be connected with the gas atmosphere in the crystal puller. In order to decide between these hypotheses, this investigation was undertaken.

For this purpose crystals were grown in a variety of atmospheres and the results correlated. As additional data in the course of the metallurgical examination, the microhardness of the crystals was measured and the results of this investigation are reported in the second part of this paper.

2. METALLOGRAPHIC EXPERIMENTS

2.1 *Crystal growth*

Crystals were grown by the CZOCHRALSKI technique[3] from a melt contained in a silica liner in a crystal puller capable of being evacuated to 5×10^{-6} mm Hg. The seed crystal was rotated at about 15 rev/min and the growth rate was about 10 cm/hr. When crystals with similar edge-dislocation densities were required, these were grown from the same seed. The range of dislocation densities covered was from 0 to 10^6 cm^{-2}. In general the crystals were grown parallel to a [111] axis.

Crystals were grown in vacuum with and without pre-baking of the liquid melt, and in a variety of gas atmospheres without pre-baking. The gases used were 10 per cent H_2 in N_2 purified by passing through a "Deoxo" catalytic purifier (Baker Platinum Ltd.) and a liquid-nitrogen trap, or pure argon 99·995 per cent (British Oxygen Gases Ltd.) either with or without a heated tungsten filament in the gas circuit. Doping with oxygen during growth was achieved by the addition of either small amounts of GeO_2 to the melt or of oxygen to the argon atmosphere. Some samples of germanium grown in oxygen atmospheres were also obtained from Mr. P. PENNING.

The resultant crystals in general contained about 10^{14} cm^{-3} shallow acceptors as determined by the Hall effect. Oxygen-doped crystals contained up to 10^{15} donors cm^{-3}. The starting material was germanium zone-refined in a N_2/H_2 mixture and having a resistivity over 50 Ω-cm at 19°C.

2.2 *Etching techniques*

The specimens $2 \times 1 \times 3$ mm were mounted with (111) faces unshielded in cold setting "Ceemar" resin (E. M. Cromwell and Co. Ltd.) to enable accurate location of the crystal for etch-pit tracing. They were ground with 15-μ alumina grinding powder (303 grade British American Optical Company) for just long enough to remove any previously etched surface. In most cases 5–10μ were removed.

The composition of the etches used is given in the Appendix. As a rule, etching was carried out in CP4[4] at 17°C for 3 min.

With CP4, two types of etch pit were generally found. These are shown in Fig. 1. Occasionally small pits (Fig. 2) were found in the lower portions of fairly rapidly cooled crystals.

Occasionally etching was carried out in CP1 at 17°C for 3 min to reveal the same type of pits on the (100) face.

The pits marked *A* shown in Fig. 1 are thought[5] to be due to dislocations. When the thickness of the crystal was decreased by successive grinding and etching (70μ being removed in each combined step), observation of these etch pits after each step showed that the dislocations responsible for them could be traced through the crystal for distances up to 300μ.

The shallow pits shown in Fig. 1 (marked *B*) and Fig. 2 could not be traced by this technique. Elimination of the grinding and reduction of the etching time to 30–60 sec reduced the step depth to 5–10 μ making it possible to examine faces that distance apart. It was found by this technique that the large shallow pits could be traced for 20–30 μ. This is approximately equal to the diameter of the etch pit and suggests that we are dealing with a spherical imperfection. The small shallow pits had an approximate diameter of 2–4 μ and could not be traced by either of these techniques.

2.3 *Results*

The following results were obtained in experiments with crystals grown under various conditions.

(a) For a crystal grown in a constant atmosphere of H_2/N_2, the concentration of large shallow pits (see *B* in Fig. 1) decreases slightly down the length (for example from 10^3 cm^{-2} at the seed end to 600 cm^{-2} near the end last grown). Smaller shallow pits (see Fig. 2) are found in the last section of the crystal to grow; when they appear, the concentration of the larger pits is rapidly reduced to zero. The number of smaller shallow pits increases rapidly from the point at which they first appear until it reaches an approximately constant level usually between 10^5 and 10^6 cm^{-2} when the larger pits have vanished.

(b) Annealing at 600°C in vacuum or pure H_2/N_2 mixtures has the effect of decreasing the number and increasing the size of the small pits.

(c) Low shallow pit densities were found in crystals grown in purified N_2/H_2 atmospheres.

(d) Changing the atmosphere in the puller during growth may cause a fairly rapid change in shallow-pit concentration. For example, a crystal grown in argon (99·995 per cent) contained about 1×10^4 cm^{-2} shallow pits. Switching on the tungsten filament in the gas circuit (presumably leading to the removal of oxygen from the gas) caused a decrease to about 200 cm^{-2} in the course of 30 min growth [compare with (a)].

(e) Starting with pit-free single-crystal material, regrowing in argon (99·995 per cent), not purified over the tungsten filament, caused the reappearance of shallow pits. (With purified gas the pits did not reappear.)

(f) Addition of oxygen to the gas stream or of

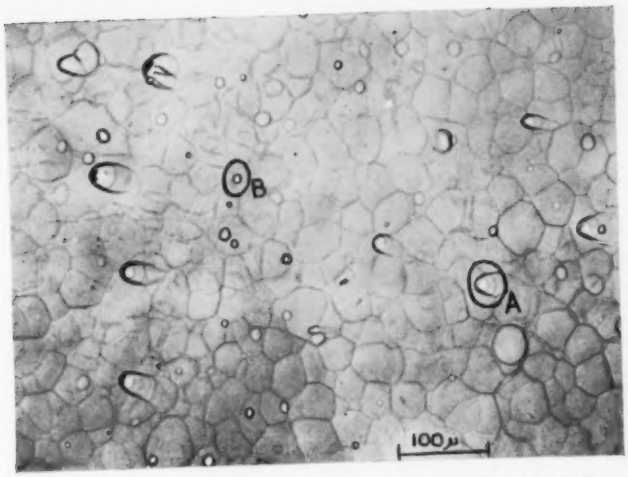


FIG. 1. Edge-dislocation etch pits (marked *A*) and shallow etch pits (marked *B*). (Magnification ×140.)

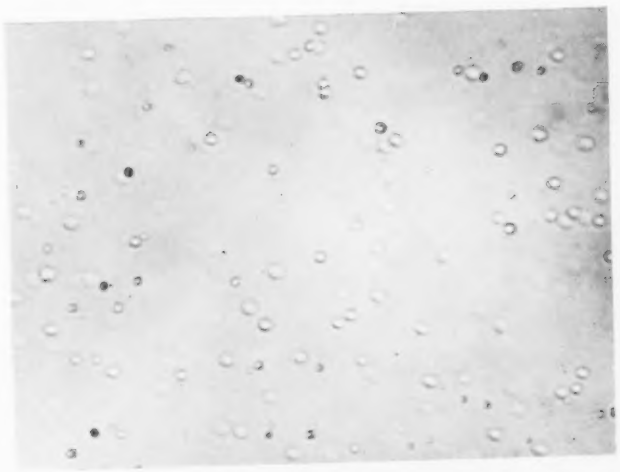


FIG. 2. Small shallow etch pits (magnification ×420).

[*facing p.* 106

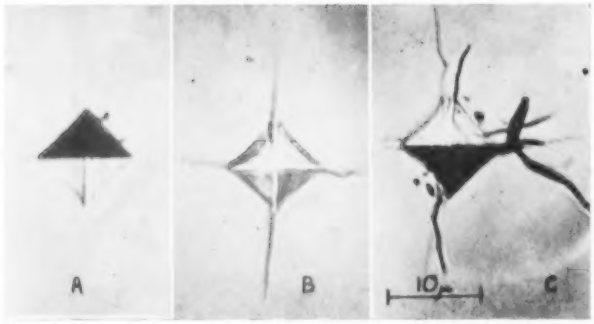


FIG. 5. Microhardness impressions 20-g load. Magnification ×1400. (a) Uncracked, (b) slightly cracked, (c) badly cracked.

GeO_2 to the melt produced an increase in shallow-pit density.

(g) Baking the liquid melt under vacuum before growth was commenced reduced the shallow-pit density. Fig. 3 shows the effect of various lengths of bake at 1100°C on the shallow-pit density for crystals grown in vacuum from the same batch of starting material.

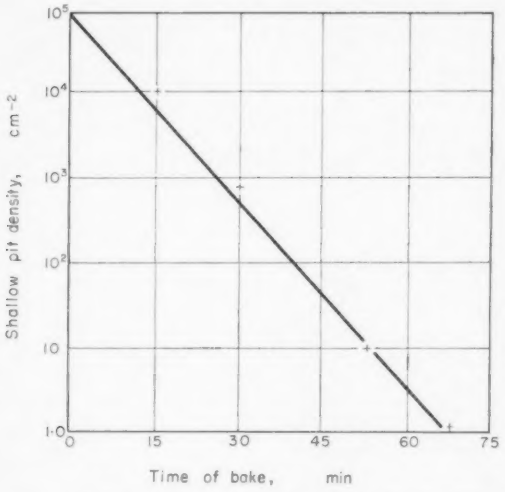


FIG. 3. The effect of baking the liquid germanium at 1100°C under vacuum before growth.

(h) The observations (c)–(g) suggest a correlation between shallow pits and oxygen. In order to test the existence of such a correlation, the oxygen content of a variety of crystals was investigated by two methods. The first was vacuum fusion applying a correction for surface oxygen. This had a limit of detection of about 2×10^{16} cm^{-3} and a probable error of about 2×10^{16} cm^{-3}. In the second method, infra-red absorption[6] was used as a measure of oxygen content.

In this case the limit of detection was about 3×10^{15} cm^{-3}, depending on sample thickness.

Oxygen content as a function of shallow pit density is shown in Fig. 4. The limits shown in this figure indicate the experimental scatter. The closed symbols have no significance except to indicate the method used. Where no lower limit is shown, no oxygen was detected and the limit of detection corresponds to the upper limit shown. In view of the experimental errors in the determinations of both quantities plotted, the correlation appears to be well established.

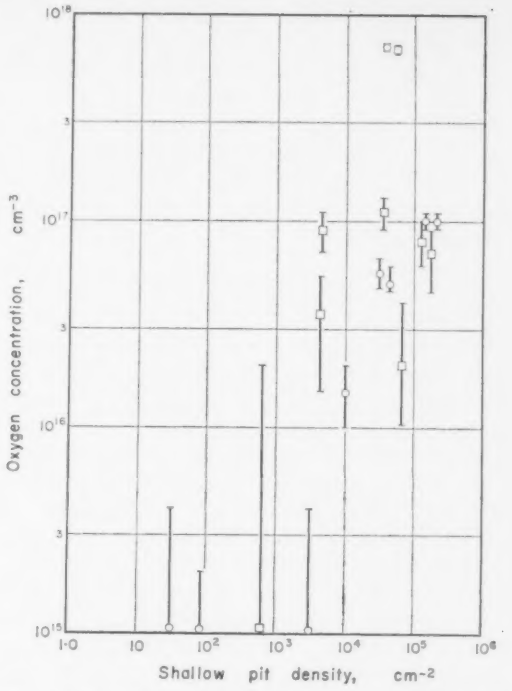


FIG. 4. The relation between oxygen and shallow-pit concentrations (□, vacuum fusion; ○, optical absorption). The symbols only indicate the method used.

2.4 *The nature of the shallow pits*

From the results described in Sections 2.2 and 2.3 it can be deduced that the shallow pits are associated with spherical (rather than linear) defects with an apparent diameter less than 30 μ, the smaller pits appear to cluster to form the larger and the shallow pits are associated in some way with an impurity taken up from the gas phase or melt, presumably oxygen.

TWEET assumes that the shallow pits are associated with vacancy aggregates in dislocation-free sections of the crystal. We found large concentrations of shallow pits in crystals having dislocation densities up to 10^6 cm^{-2}. Since the vacancies would precipitate on the dislocations under such conditions, the pits cannot be due to pure vacancy clusters. TWEET's model also fails to account for the correlation with impurities (oxygen) found.

It would be possible to modify TWEET's model to encompass the impurity effects by assuming a composite cluster formed by vacancies clustering on the oxygen. Another possibility would be to assume oxygen clusters on vacancies. Finally there would be the further possibility of clusters of oxygen on some other nucleus. Further work is required to distinguish between these three possible models.

3. HARDNESS TESTING

3.1 *Technique*

A Vickers Microhardness Tester was used for all the hardness tests. This is essentially a diamond indentor that is applied to the test piece under a given load, at a known rate for a set time.

For germanium the best conditions were found to be the use of a 20- or 30-g load applied for 15 sec at a rate not exceeding 1–2 μ per sec. From the length of the diagonals of the resulting impression, the Vickers Pyramid Numeral (VPN) can be derived from conventional formulae. Since most of the diagonals were between 8 and 12μ long, the standard modification to the calculated result had to be made to allow for the slight ridge at the apex of the indentor.

Prior to testing, the samples, which were approximately 1 cm dia. and 3 mm thick, were mounted in cold-setting resin blocks (1 in. dia. to fit the chuck on the Micro Hardness Tester).

Since mechanically polished surfaces were found to be highly strained and to cause a great deal of conchoidal fracturing on the application of the indentor, the surfaces of the test pieces were prepared by grinding on 15-μ alumina and etching in hot CP4 for 3–4 min to give an "orange peel" free surface and to remove about 300 μ of germanium. If less than 300 μ were removed, consistent results were difficult to obtain owing to the high incidence of cracking from the corners of the impressions. After removal of 300 μ, the residual cracking could be reduced to a minimum by rotating the microscope stage through 90°, making impressions every 15° and then selecting the most favourable orientation.

When the crystals were aligned in the optimum direction, 5–10 per cent of impressions could be obtained without visible signs of cracking, a further 85 per cent had very fine cracks; the diagonals of these were, in all cases, within $\pm$ 4 per cent of the uncracked impressions. About 5 per cent of the impressions were heavily cracked and chipped, the results from these being discounted. Typical impressions are shown in Fig. 5.

All the VPN values used in this paper were obtained from the average of nine or more impressions made on (111) faces (orientated correctly to $\pm$ 1°). The scatter in all cases was less than $\pm$ 4 per cent. In the majority of samples tested, the hardness was sensibly constant over the cross section but no tests were carried out in the outmost millimetre of sample.

3.2 *The effect of etch pits on hardness*

As would be expected the effect of increasing the concentration of edge dislocations (Fig. 6) is

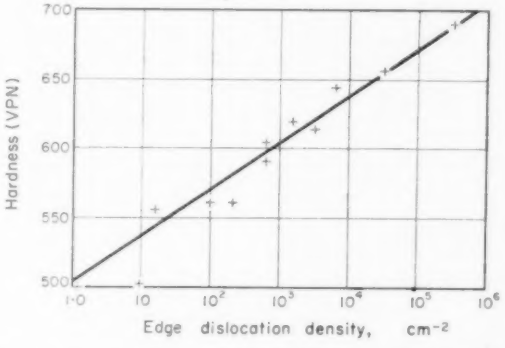


FIG. 6. The effect of edge dislocations on hardness (shallow-etch-pit density constant at 1×10^3 cm^{-2}).

that of increasing hardness, in this case from 500 VPN for dislocation free material to 700 VPN for material with 10^6 edge dislocations cm^{-2}. Since the shallow pits have a similar effect (Fig. 7) it must be concluded that the defect causing these pits also introduces strain into the lattice though to a smaller extent than the edge dislocations.

4. CONCLUSIONS

The work described in this paper has shown that the shallow etch pits found on etching germanium are due to globular defects introduced by contamination, probably by oxygen.

Both the shallow pits and edge dislocations produce an increase in hardness which, in the case of edge dislocations, reaches about 700 VPN at 5×10^6 edge dislocations per cm^2.

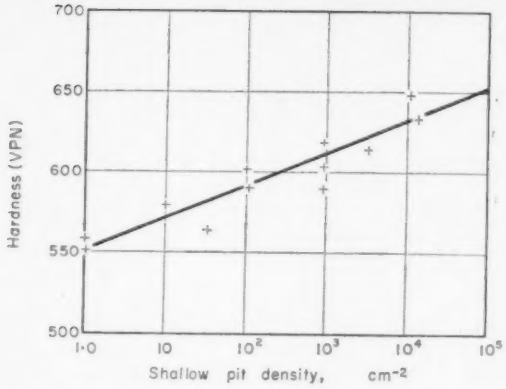


FIG. 7. The effect of shallow-etch-pit density on hardness (edge dislocation density constant at 1×10^3 cm^{-2}).

Acknowledgements—The authors thank Dr. F. A. KRÖGER for considerable help with the interpretation of the results. The practical and theoretical help of many members of the Solid State Physics Division is gratefully acknowledged. The optical measurements were carried out by Mr. K. E. G. PITT and Dr. C. HAAS, vacuum-fusion determinations by Mr. E. J. MILLETT and the crystals grown in known oxygen pressures were supplied by Mr. P. PENNING.

REFERENCES

1. F. D. ROSI, *RCA Review* **19,** 349 (1958).
2. A. G. TWEET, *J. Appl. Phys.* **30,** 2002 (1959).
3. J. CZOCHRALSKI, *Z. Phys. Chem.* **92,** 219 (1918).
4. J. R. HAYNES and W. SHOCKLEY, *Phys. Rev.* **81,** 835 (1951).
5. F. L. VOGEL, W. G. PFANN, H. E. CAREY and C. E. THOMAS, *Phys. Rev.* **90,** 489 (1953).
6. J. BLOEM, C. HAAS and P. PENNING, *J. Phys. Chem. Solids* **12,** 22 (1959).

APPENDIX

The compositions of the etches used:

CP4	HNO_5	10 parts
	HF	6 parts
	Acetic acid (glacial)	6 parts
	Bromine	2 drops in 50 ml
CP1	HNO_3	1 part
	HF	1 part
	Acetic acid glacial	2 parts
	Iodine	40 mg in 50 ml

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 110–118. Printed in Great Britain

AN ASTABLE FOUR-REGION MULTIVIBRATOR*

G. STRULL* and W. M. KAUFMAN†

Westinghouse Electric Corporation

(*Received* 4 *November* 1960; *in revised form* 21 *March* 1961)

Abstract—A molecular engineered astable multivibrator has been developed which utilizes a four-region semiconductor structure. This structure with a power supply and load acts as a relaxation oscillator. The operation of this structure has been analyzed using a linearized characteristic to provide some insight into the effect of various parameters. Certain temperature effects have been found detrimental to the operation of the multivibrator. By the extension of this structure to more than four regions, temperature stability may be obtained. The results obtained with a five-region structure are also presented.

Résumé—Un multivibrateur astable de construction moléculaire employant une structure semiconductrice à quatre régions a été développé. En joignant cette structure à une charge et une alimentation, elle devient un oscillateur à relâchement.

Le fonctionnement de cette structure a été analysé en employant une caractéristique linéaire pour fournir un aperçu des effets des divers paramètres. On a trouvé que certains effets de température sont nuisibles au fonctionnement du multivibrateur. En étendant cette structure à plus de quatre régions, on peut obtenir une plus grande stabilité en fonction de la température. Les résultats obtenus avec une structure à cinq régions sont aussi presentés.

Zusammenfassung—Ein astabiler Multivibrator wurde unter Verwendung eines Vierzonen-Halbleiters entwickelt. Das Gerät wirkt bei Kraftstrombetrieb und Belastung als Relaxationsoszillator.

Seine Wirkungsweise wurde mittels einer linearisierten Kennlinie untersucht, um Einblick in den Effekt verschiedener Parameter zu gewinnen. Es ergab sich, dass gewisse Temperatureffekte die Leistung des Multivibrators beeinträchtigen. Es ist möglich, dass bei Erweiterung des Gerätes auf mehr als vier Zonen Temperaturbeständigkeit erzielt werden kann. Die mit einem Fünzonengerät erhaltenen Ergebnisse werden angegeben.

INTRODUCTION

It is possible to obtain an astable multivibrator with a four-region semiconductor structure, power supply and load. The shunt capacitance generally required in the classical (gas tube) multivibrator of this type is an inherent part of the semiconductor structure. The four-region structure required for multivibrator operation is related to the four-region n–p–n–p switch. However, the multivibrator structure has different requirements on the width of the negative-resistance region.

STRUCTURES

In Fig. 1 is shown the characteristics of the n–p–n–p structure necessary for multivibrator operation. As can be observed, the breakover current should be low and the minimum sustaining current high. For the multivibrator to operate successfully, the load line must lie in the negative-resistance region of the V–I characteristics of the four-region structure. If the negative-resistance region is narrow, as would be desired for a good switch, then high external resistances are required and the necessary applied voltage would have to be large compared to the breakover voltage of the structure. If, on the other hand, the minimum sustaining current is high, the load line can have a shallow slope, and the applied voltage required for oscillation approaches the breakover voltage of the structure.

ANALYSIS

A linearized analysis was performed for this structure in order to obtain some insight into the

* Semiconductor Division, Baltimore, Maryland.

† Research Laboratory, Pittsburgh, Pennyslvania.

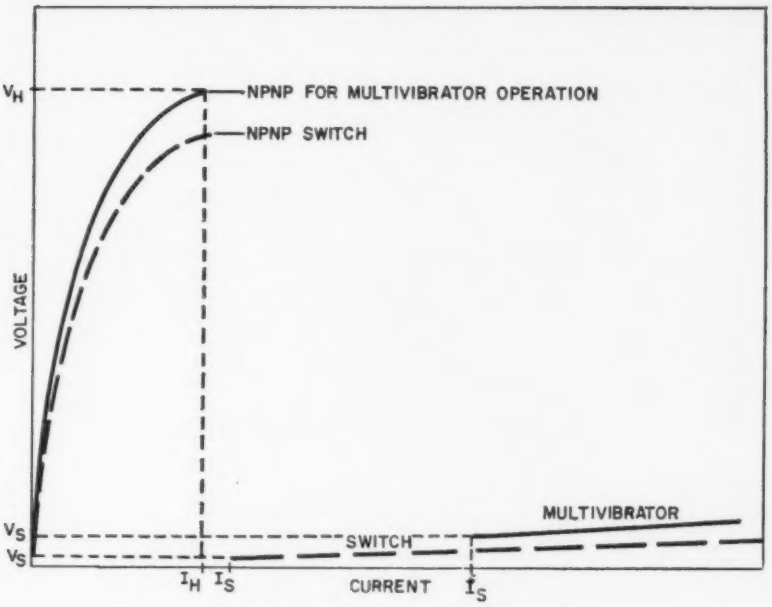


FIG. 1. n–p–n–p characteristic for multivibrator operation.

importance of the various design parameters with respect to frequency of oscillation. The analysis that was performed was the classical analysis of a relaxation oscillator. It is not suggested that the results of this analysis will provide exact information concerning the voltage across the device as a function of time, but the effects of such parameters as the holding voltage and current, the sustaining voltage and current, the source voltage, the source impedance and the effective junction capacitance on oscillation frequency should be demonstrated.

Although the static voltage–current (v–i) characteristic of this device is highly nonlinear (see Fig. 1), one way of obtaining a quantitative picture of the operation is to assume piece-wise linearity of this characteristic. Fig. 2(a) shows how this can be done and Fig. 2(b) contains the equivalent circuit of the linearized device. The junction capacitance C is actually voltage dependent, but for the present purposes will be assumed constant in order to maintain linearity. This assumption causes the results to be somewhat approximate and is one of the many approximations made in order to obtain a simple form for an estimate of the maximum frequency of oscillation. In the equivalent circuit of Fig. 2(a), R_L includes the series summation of load resistance and internal resistance of the voltage source. The series resistance of the p–n–p–n structure is included in the linearized v–i characteristic. S_1 and S_2 are switches which correspond to the switching between the two branches of the v–i characteristic. Switch S is the switch that energizes the circuit. The oscillations proceed in the following fashion. Assume capacitor C is uncharged prior to energizing the circuit. With the closing of S, C begins to charge (exponentially in time) and the capacitor voltage rises toward $R_1E/(R_1+R_L)$. (S_1 is closed during this time.) However, if $V_H < R_1E/(R_1+R_L)$, then the system will switch to the other branch of the v–i curve (S_1 opens, S_2 closes) when voltage v reaches V_H. Immediately upon S_2 closing, the capacitor voltage begins to decrease toward $R_2E/(R_2+R_L)$. If $V_S < R_2E/(R_2+R_L)$, then the system will switch back to the initial state (S_1 closed, S_2 open) when v reaches V_S. Thus cyclic operation is sustained.

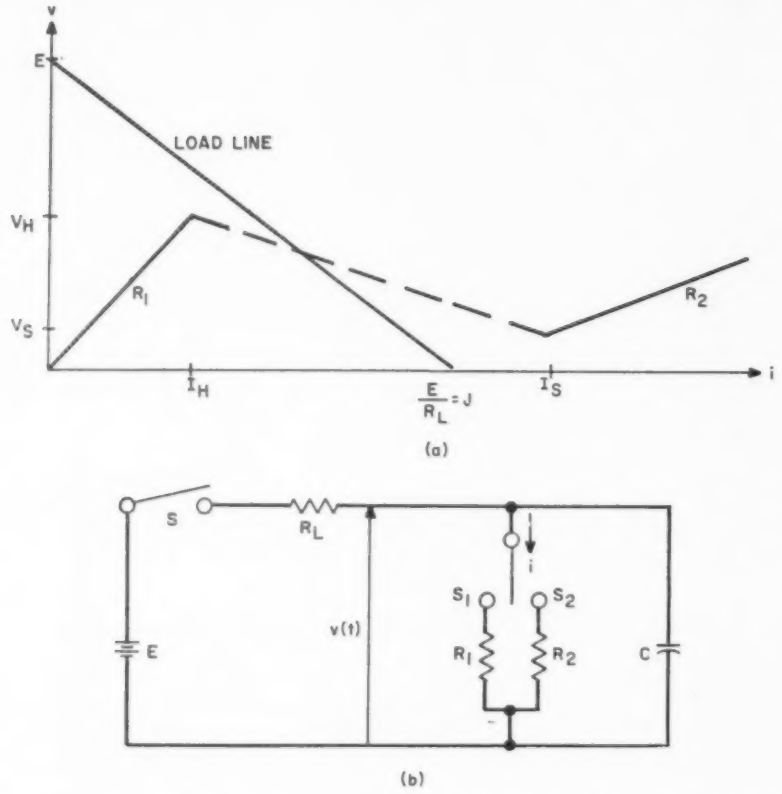


FIG. 2(a). Linearized characteristics for the four-region multivibrator. (b). Equivalent circuit for analysis of the multivibrator.

Let State I be defined as that for which S_1 is closed and State II be that for which S_2 is closed. Then the period of oscillation P is the sum of the time, t^*, spent in State I plus the time, t^{**}, spent in State II during one cycle of operation. (Details of the mathematical manipulations are given in Appendix A.)

$$P = t^* + t^{**} \tag{1}$$

$$t^* = \frac{R_1 R_L C}{R_1 + R_L} \ln \frac{V_S - ER_1/(R_1+R_L)}{V_H - ER_1/(R_1+R_L)} \tag{2}$$

$$t^{**} = \frac{R_2 R_L C}{R_2 + R_L} \ln \frac{V_H - ER_2/(R_2+R_L)}{V_S - ER_2/(R_2+R_L)} \tag{3}$$

In a practical device,

$$V_S/E \ll 1,\ R_2/(R_L+R_2) \ll V_H/E, \text{ and } R_2/R_L \ll 1.$$

Then

$$P \simeq \frac{R_L}{R_1+R_L} R_1 C \ln \frac{1-R_L/(R_1+R_L)}{1-V_H/E-R_L/(R_1+R_L)} + R_2 C \ln \frac{V_H/E}{V_S/E-R_2/R_L} \tag{4}$$

The major concern of this analysis was to determine the minimum period of oscillation. As presented in Appendix A, the analysis indicated that a minimum period of oscillation is obtainable

when $R_L \gg R_1$. For this condition

$$P \simeq R_1 C \ln \frac{J}{J - V_H/R_1} + R_2 C \ln \left[\frac{V_H}{R_2} \frac{1}{V_S/R_2 - J} \right] \quad (5)$$

where $J = E/R_L$ is the only independent variable for a given p–n–p–n structure.

For a minimum period,

$$J \simeq \frac{R_1}{R_1 - R_2} \frac{V_S}{R_2} \left[1 - \left(\frac{R_1}{R_1 - R_2} \right)^2 \frac{V_S}{V_H} \right] \quad (6)$$

Since $V_S/R_2 = I_S$ and $V_H/R_1 = I_H$, and since (for most devices) $R_1 \gg R_2$, then

$$J = \frac{E}{R_L} \simeq I_S \left(1 - \frac{V_S}{V_H} \right) \quad (7)$$

for a minimum period.

If $R_L \gg R_1$ is to hold, then

$$E \gg \frac{I_S}{I_H}(V_H - V_S) \qquad (\text{with } R_L \gg R_1) \quad (8)$$

must be satisfied. Expression (8) is a condition that relates the circuit parameters to the device parameters in order to obtain a minimum period of oscillation.

The shortest period of oscillation possible for a given device can be found approximately by substituting (7) into (5). This yields

$$P_{\min} \simeq \frac{V_H}{I_H} C \ln \frac{1 - V_S/V_H}{1 - V_S/V_H - I_H/I_S} + 2 \frac{V_S}{I_S} C \ln \frac{V_H}{V_S} \quad (9)$$

Equation (9) indicates the effects of the various design parameters of the four-layer device upon maximum attainable frequency of oscillation. The minimum period has been found here (approximately) in terms of device parameters only.

It is apparent from equation (9) that reduction in the effective junction capacitance is one way of reducing the minimum period. This is not surprising and was intuitively expected from the beginning. However, at this point a word of caution should be expressed. If the effective capacitance is too small a completely different mode of operation can exist. This is due to the fact that the junctions possess an inductive component. The "inductance diode" is a well-known example.[1] In order to obtain some notion of the relative effects of the various design parameters, a similar linearized analysis was performed under the assumption that there is an inductance inherently associated with the negative-resistance region of the device v–i characteristic ($I_H < i < I_S$). The linear equivalent circuit that was considered is shown in Fig. 3. This equivalent was chosen

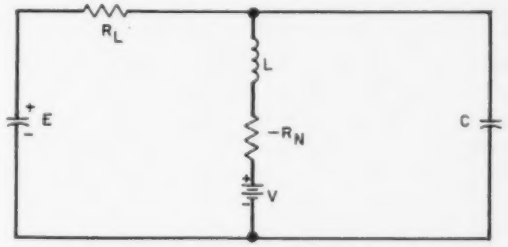


FIG. 3. Equivalent circuit for the linearized characteristics in the negative-resistance region.

instead of one used for the tunnel diode because it has been found experimentally that many four-layer switches can be biased stably in their negative-resistance regions when driven from a current source. Appendix B shows that relaxation oscillations will occur as long as

$$C > \frac{L}{R_N^2} \left[2 - \frac{R_N}{R_L} + 2 \sqrt{\left(1 - \frac{R_N}{R_L} \right)} \right] \quad (10)$$

or, for $R_L \gg R_N$, this becomes

$$C > \frac{4L}{R_N^2} \quad (10')$$

If the inequality (10) is not met, the device will either oscillate about an operating point in the negative-resistance region of the v–i characteristic or will remain stably in the negative-resistance region. The condition for stability in the negative-resistance region is

$$C < \frac{L}{R_L R_N} \quad (11)$$

as shown in Appendix B. Between these two conditions (10) and (11) is the state in which sinusoidal oscillation takes place.

Since all the above analysis was performed on idealized, linearized lumped-parameter models the results are meant to be used as a guide for device design. Equation (9) shows how the various device design parameters interact to affect the minimum period of oscillation, and equation (10) indicates the relationship between device capacitance and device inductance in order to maintain relaxation oscillations.

FABRICATION

The structure utilized for the multivibrator is shown in Fig. 4. The fabrication procedure is similar to that used for *n–p–n–p* switching structures, with the major distinction the fact that, in a good switch, steps are taken to reduce the minimum sustaining current by providing for a high alpha at low currents, while the multivibrator requires generally higher minimum sustaining currents for paractical operation. It is extremely important to note that the high minimum sustaining current is a practical requirement in connection with the range of loads and applied voltages for operation, but even units with low minimum sustaining currents can function as described.

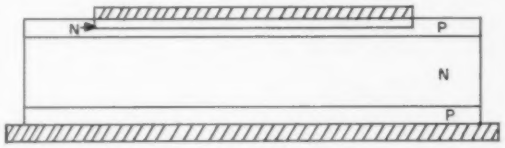


FIG. 4. Multivibrator structure.

The multivibrators have been fabricated in both germanium and silicon. The silicon multivibrators were made both using single diffusion followed by alloy fusion and by an all-diffusion process. In

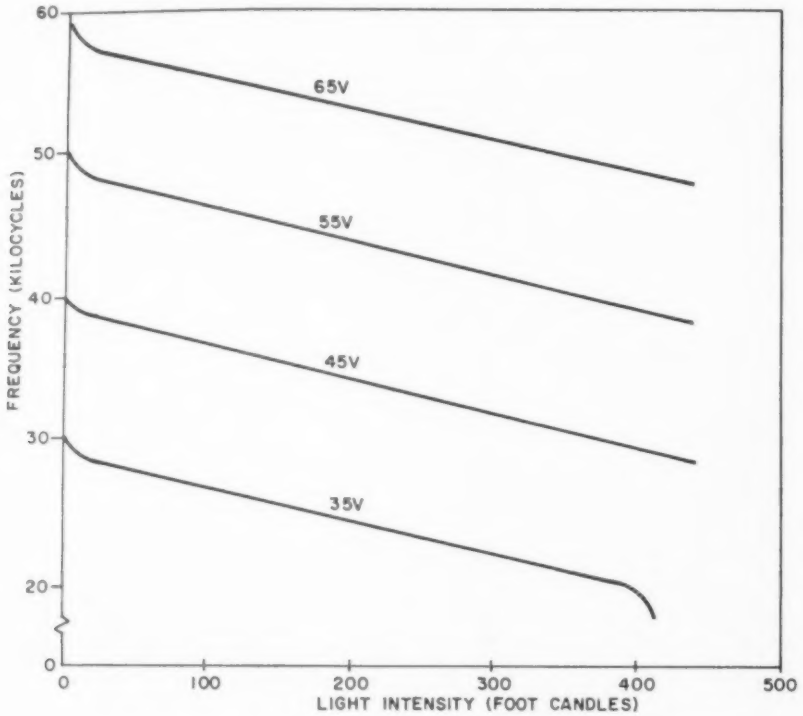


FIG. 5. Frequency vs. light intensity for a four-region multivibrator.

germanium, most of the units were made by a diffusion followed by an alloy fusion for the fourth region. In addition, the *p–n–p–m* structure[2] with its very high obtainable minimum sustaining current was also investigated in germanium.

FIVE-REGION STRUCTURES

The behavior of this multivibrator with light or temperature is quite interesting. The *v–i* characteristic before breakover moves in the direction of increasing current with increasing light or temperature, thus changing the width of the negative-resistance region and thereby changing the frequency of oscillation of the structure. The variation of frequency with light is shown in Fig. 5. The variation of frequency with temperature for a typical multivibrator is shown in Fig. 6. If an additional region is added to form a five-

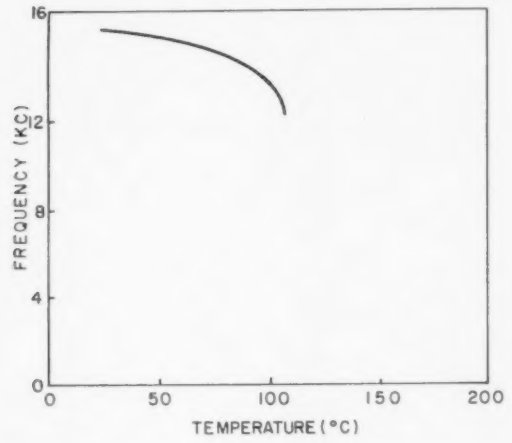


FIG. 6. Frequency vs. temperature for a four-region multivibrator.

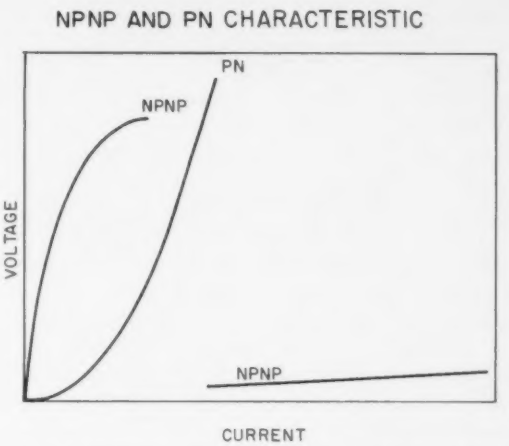


FIG. 7. Characteristics needed for composite five-region structure.

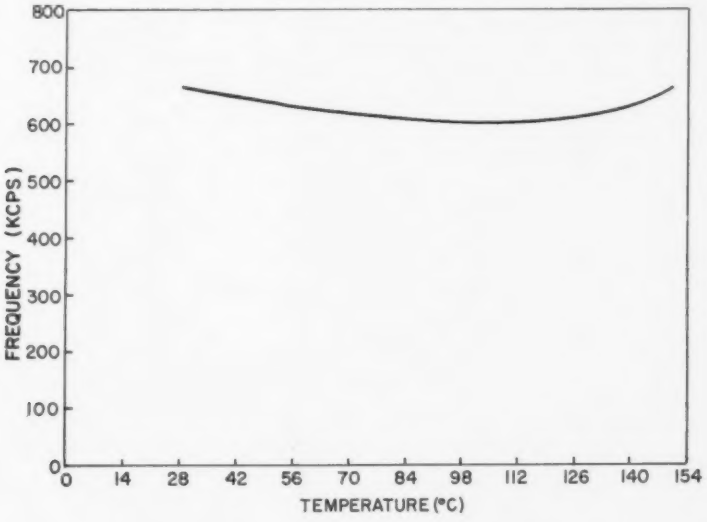


FIG. 8. Frequency vs. temperature of a five-region structure.

region *n–p–n–p–n*, it is possible to compensate for this temperature variation to a large extent. The addition of the fifth region may be considered, under proper conditions, similar to putting a diode in the negative-resistance region of the bistable structure. The diode saturation characteristic appears as shown in Fig. 7. The behavior of a structure of this type with temperature is shown in Fig. 8. The vast improvement in temperature stability between Fig. 6 and Fig. 7 is clearly evident. This occurs because, as the four-layer structure tends to increase current with increasing temperature, the diode characteristic moves in such a direction as to compensate for this, thereby retaining constant frequency of oscillation over a wide temperature range.

CONCLUSION

Structures of the type of the four-region multivibrator and particularly the five-region temperature compensated multivibrator perform electronic circuit functions without a one-to-one correspondence between some portion of the functional electronic block and a circuit performing the same task. An exact analysis of such a structure is not readily obtained. The linear analysis, coupled with experimental work, is sufficient to provide a reasonable design base for such structures.

APPENDIX A

For the equivalent circuit of Fig. 2(b), the expressions for $v(t)$ in the two states are:

State I

$$v(t) = E\frac{R_1}{R_1+R_L} - \left(E\frac{R_1}{R_1+R_L} - V_S\right)\exp(-t/\tau_{p1}) \tag{A.1}$$

$$\tau_{p1} = \frac{R_1R_L}{R_1+R_L}C$$

$t = 0$ when S_1 is closed.

State II

$$v(t) = E\frac{R_2}{R_2+R_L} - \left(E\frac{R_2}{R_2+R_L} - V_H\right)\exp(-t/\tau_{p2}) \tag{A.2}$$

$$\tau_{p2} = \frac{R_2R_L}{R_2+R_L}C$$

$t = 0$ when S_2 is closed.

Let t^* = time in State I, i.e. $v(t^*) = V_H$, and t^{**} = time in State II, i.e. $v(t^{**}) = V_S$. Then P, the period of oscillation is

$$P = t^* + t^{**} \tag{1) or (A.3}$$

$$t^* = \tau_{p1}\ln\frac{V_S - ER_1/(R_1+R_L)}{V_H - ER_1/(R_1+R_L)} \tag{2) or (A.4}$$

$$t^{**} = \tau_{p2}\ln\frac{V_H - ER_2/(R_2+R_L)}{V_S - ER_2/(R_2+R_L)} \tag{3) or (A.5}$$

Equation (4) or (A.6) is obtained by making the assumptions

$$V_S/E \ll 1,\ R_2/(R_L+R_2) \ll V_H/E, \text{ and } R_2/R_L \ll 1.$$

$$P \simeq \frac{R_L}{R_1+R_L}R_1C\ln\frac{1-R_L/(R_1+R_L)}{1-V_H/E-R_L/(R_1+R_L)} + R_2C\ln\frac{V_H/E}{V_S/E-R_2/R_L} \tag{4) or (A.6}$$

Of particular interest is the shortest period of oscillation that can be obtained with a given device. For a given device only E and R_L are variable. Thus $\partial P/dE = 0$ and $\partial P/\partial R_L = 0$ must be satisfied for a minimum in P. These expressions are very complicated and were not solved in general. Instead, the following two cases were considered: Case I, $R_L \ll R_1$, and Case II, $R_L \gg R_1$.

Case I

$$P \simeq R_LC\ln\frac{1}{1-V_H/E} + R_2C\ln\frac{V_H}{V_S-R_2E/R_L} \tag{A.7}$$

Let $E/R_L = J$, then

$$P(J,E) = \frac{CV_H}{J}\frac{E}{V_H}\ln\frac{1}{1-V_H/E} + R_2C\ln\frac{V_H}{V_S-R_2J} \tag{A.8}$$

The second term t^{**} is no longer a function of E. For a fixed value of J, the first term t^* decreases for $V_H/E \to 0$. In fact, at the limit,

$$\frac{V_H}{E}\overset{\lim}{\to}0\frac{1}{V_H/E}\ln\frac{1}{1-V_H/E} = 1 \tag{A.9}$$

Thus, for any value of J, P is reduced by increasing E. However, for fixed J, as E increases R_L increases; therefore, this runs contrary to the condition $R_L \ll R_1$. Therefore, the minimum value of P is probably obtainable in Case II where $R_L \gg R_1$.

Case II

$$P \simeq R_1 C \ln \frac{R_1 E/R_L}{R_1 E/R_L - V_H} + R_2 C \ln \frac{V_H}{V_S - R_2 E/R_L} \tag{A.10}$$

Equation (A.10) is obtained from equation (A.6) by assuming $R_L \gg R_1$. Again let $E/R_L = J$, then equation (5) is obtained.

$$P \simeq R_1 C \ln \frac{J}{J - V_H/R_1} + R_2 C \ln \left[\frac{V_H}{R_2} \frac{1}{V_S/R_2 - J} \right] \tag{5 or A.11}$$

In this case P is a function of only one variable, J. Setting $dP/dJ = 0$ and solving for J yields the value of J for a minimum P.

$$\frac{dP}{dJ} = R_1 C \left(\frac{1}{J} - \frac{1}{J - I_H} \right) + R_2 C \frac{1}{I_S - J} \tag{A.12}$$

Therefore, for $dP/dJ = 0$,

$$J = \frac{V_H}{2R_2} \left\{ -\left(1 - \frac{R_2}{R_1}\right) + \sqrt{\left[\left(1 - \frac{R_2}{R_1}\right)^2 + \frac{4V_S}{V_H} \right]} \right\} \tag{A.13}$$

Since $R_2/R_1 \ll 1$ and $V_S/V_H \ll 1$ in a typical device, the binomial expansion for the square root will be used and all terms containing

$$\left(\frac{V_S}{V_H} \right)^k, \quad k > 2,$$

will be neglected. Then the expression for J simplifies to

$$J \simeq \frac{R_1}{R_1 - R_2} \frac{V_S}{R_2} \left[1 - \left(\frac{R_1}{R_1 - R_2} \right)^2 \frac{V_S}{V_H} \right] \tag{6 or A.14}$$

or

$$J \simeq I_S \left(1 - \frac{V_S}{V_H} \right), \qquad \frac{R_1}{R_1 - R_2} \simeq 1 \tag{7 or A.15}$$

for a minimum period of oscillation.

APPENDIX B

For the equivalent circuit of Fig. 3, the differential equation for current i through the $L - R_N$ branch is

$$0 = R_L LC \frac{d^2 i}{dt^2} + (L - R_L R_N C) \frac{di}{dt} + (R_L - R_N) i \tag{B.1}$$

The mode of oscillation can be determined from the coefficients of this differential equation. For example, the circuit will not oscillate if all the coefficients in (B.1) have the same sign. Then, if $R_L > R_N$ and $L > R_L R_N C$, the circuit is stable in the negative-resistance region; i.e., for stability,

$$R_N < R_L < \frac{L}{R_N C} \tag{B.2}$$

For relaxation oscillation, it is necessary that $R_L > R_N$, otherwise the device will operate stably in either of its two positive resistance states. Therefore, $R_L > R_N$ will always be assumed. As mentioned above, if $L > R_L R_N C$ the circuit is stable. Therefore, for some type of oscillation $L < R_L R_N C$ is required.

Consider the two roots of the characteristic equation for (B.1).

$$M_{1,2} = \frac{R_1 R_N C - L \pm \sqrt{[(R_L R_N C - L)^2 - 4 R_L LC (R_L - R_N)]}}{2 R_L LC} \tag{B.3}$$

with $R_L > R_N$ and $L < R_L R_N C$. If $4R_L LC(R_L - R_N) > > (R_L R_N C - L)^2$, then the solution for $i(t)$ is an exponentially growing sinusoidal oscillation with angular frequency

$$\omega_0 = \sqrt{\left[R_L - R_N - \tfrac{1}{4} \left(\frac{1}{R_L C} - \frac{R_N}{L} \right)^2 \right]} \tag{B.4}$$

If $4R_L LC(R_L - R_N) < (R_L R_N C - L)^2$, then $i(t)$ is the sum of two growing exponentials. This inequality can be rewritten in the form, for two growing exponentials,

$$C > \frac{L}{R_N^2} \left[2 - \frac{R_N}{R_L} + 2 \sqrt{\left(1 - \frac{R_N}{R_L} \right)} \right] \tag{B.5}$$

after some simple mathematical manipulations.

The exponentially growing sinusoidal solution is a distinctly different mode of operation from the two growing exponentials solution. With the linear approximation there is nothing to prevent either of these solutions from growing indefinitely; however, because of the nonlinear character of the device, the sinusoidal solution eventually stops growing, while the exponential solution permits relaxation oscillations. The sinusoidal solution stops growing when the amplitude of the voltage swings increases to the point where the average effective negative resistance $\bar{R}_N$ becomes small enough to satisfy the equation

$$\bar{R}_N = \frac{L}{R_L C} \tag{B.6}$$

BRUNETTI[3] expressed the average effective negative resistance as a function of voltage swing for sinusoidal voltages is terms of energy considerations.

$$\bar{R}_N = \frac{V^2}{2(\text{power supplied by the nonlinear device})} \tag{B.7}$$

where V is the maximum value of the sinusoidal voltage across the device. BRUNETTI's work shows how this can be used to predict the amplitude of sinusoidal oscillations.

REFERENCES

1. J. NISHIZAWA and Y. WATANABE, *Semiconductor Inductance Diode*, International Solid-State Circuits Conference, University of Pennsylvania, Pennsylvania, (1960).
2. J. PHILIPS and H. C. CHANG, *I.R.E. Trans.* **ED-5,** 13 (1958).
3. C. BRUNETTI, *Proc. I.R.E.* **25,** 1595, 1616 (1937).

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 119–122. Printed in Great Britain

TURN-OFF GAIN IN *p–n–p–n* TRIODES*

J. M. GOLDEY, I. M. MACKINTOSH and I. M. ROSS

Bell Telephone Laboratories, Inc., Murray Hill, New Jersey

(*Received* 13 *March* 1961)

Abstract—It has long been recognized that large turn-on gain can be achieved in *p–n–p–n* triodes when driven through the base. On the other hand, it has been recognized only recently that these devices can also be turned off through the base with large gain. In this paper we shall discuss in detail the mechanism of base-current turn off and show how large turn-off gain can be achieved. Several structures in which large turn-off gain can be attained will be described. Electrical characteristics of experimental models which show gains approaching 40 at collector currents of about 100 mA will be presented.

Résumé—On a reconnu depuis assez longtemps que de grands gains de déclenchement peuvent être obtenus en employant des triodes *p–n–p–n* controllées par la base. D'un autre côté, on n'a reconnu que récemment que ces éléments peuvent être coupés par la base avec un grand gain. Dans cet article, on décrira en détail le mécanisme de coupure du courant de base et on démontrera comment un grand gain de coupure pourra êtra achevé. Plusieurs structures dans lesquelles un grand gain de coupure peut être atteint seront décrites. Les caractéristiques électriques des modèles expérimentaux qui produisent des gains approchant 40 à des courants de collecteur d'environ 100 mA seront présentées.

Zusammenfassung—Es ist lange bekannt, dass bei *p–n–p–n*-Trioden beim Einschalten ein grosse Verstärkung erzielt werden kann, wenn der Strom an die Basis gelegt wird. Man hat aber erst in jüngster Zeit erkannt, dass auch beim Abschalten durch die Basis eine grosse Verstärkung zu erzielen ist. In der vorliegenden Arbeit wird der Mechanismus des Abschaltens mittels Basisstrom diskutiert und gezeigt, dass eine hohe Verstärkung möglich ist. Wir beschreiben ferner mehrere Bauelemente, die hohe Abschaltverstärkungen ergeben und geben Einzelheiten über Versuchsmodelle, die bei Kollektorströmen von etwa 100 mA Verstärkungen bis zu 40 liefern.

INTRODUCTION

THE properties and characteristics of three-terminal *p–n–p–n* switches have been discussed by MACKINTOSH[1] and by MUELLER and HILIBRAND[2] in some detail. These authors have described the static characteristics, the turn-on mechanism and other details of operation. They have not described the mechanism of turn-off by withdrawal of base current in any detail, however. In this paper this mechanism will be discussed. Several structures which provide base control with large current gain in turning off as well as turning on will be described and experimental results on laboratory models will be given.

*This material was originally presented at the IRE–AIEE Solid State Device Research Conference, Pittsburgh, Pennsylvania, June 1960.

MECHANISM

A *p–n–p–n* diode or triode is in many ways equivalent to a conjugate pair of transistors connected with collector junctions in parallel in the so-called flip-flop configuration.[3] When the device is in the low-impedance state, each of the component transistors is in the saturated condition. The drive current required to saturate each transistor is supplied by the collector or output current, or a portion thereof, of the other transistor. In Fig. 1 a schematic of the *p–n–p–n* triode and the current flow directions are shown. α_n is the d.c. current multiplication ratio of the *p–n–p* and α_p that of the *n–p–n*. The drive current, i.e. the net majority-carrier current into the base, required to maintain the *n–p–n* just in saturation is $(1-\alpha_p)I_E$. The maximum drive current available,

with the base lead open-circuited, is the collector current, $\alpha_n I_C$, of the p–n–p. Thus, for the device to be on with the base lead open-circuited, ($I_C = I_E$), we have

$$\alpha_n + \alpha_p \geqslant 1 \tag{1}$$

In order to return the p–n–p–n to its high impedance state, the drive current must be reduced

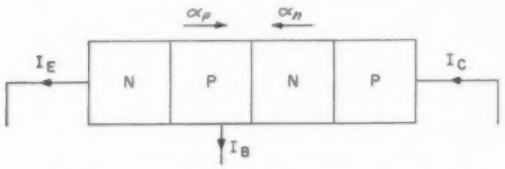


FIG. 1. Schematic representation of p–n–p–n triode showing directions of current flow and current multiplication factors. α_n is the d.c. current gain of the p–n–p and α_p that of the n–p–n.

below the value required for saturation. This current can be most easily calculated by considering one of the component transistors. Since the two component transistors have a common collector junction, when one comes out of saturation the other does also, and consequently the p–n–p–n goes off. When majority carriers are removed from the base of the n–p–n transistor through the base lead, the maximum drive current to the n–p–n becomes $\alpha_n I_C - I_B$. Thus, the n–p–n will be out of saturation when

$$\alpha_n I_C - I_B < (1-\alpha_p) I_E \tag{2}$$

or, since

$$I_C = I_E + I_B \tag{3}$$

then

$$(\alpha_n + \alpha_p - 1) I_C < \alpha_p I_B \tag{4}$$

As in the case of a saturated three-region transistor, the collector current is determined by the external circuit. Consequently changes in base current have little effect on the collector current while the device is on. Hence the α's of interest in equation (4) are those corresponding to the "on" collector current and vary but little until the switch comes out of saturation. Once turn-off is achieved the collector current drops to near zero and correspondingly the α's drop. Thus the device remains off when the base is open-circuited.

If we define the turn-off gain as the ratio of current flowing in the load, when the switch is on, to the base current required to turn the switch off, then the turn-off gain is given from equation (4) by

$$\frac{I_C}{I_B} = \frac{\alpha_p}{(\alpha_n + \alpha_p - 1)} \tag{5}$$

Thus, high turn-off gain can be achieved in p–n–p–n triodes if the device is designed in such a way that, when on, the α sum just exceeds unity and the α of the transistor to which the base lead is attached (α_p here) is high. Physically this means that the switch (or the component transistors) is just in saturation. The excess stored charge is minimized thereby minimizing the base current required for turn off.

In order to minimize the drive to turn on, it is again necessary for the α of externally driven transistor (α_p here) to approach unity. Thus maximum turn-on and turn-off gains are both achieved when the α of the base to which external connection is made (α_p here) is high, and the other α (α_n here) is low.

STRUCTURES

In this section we shall describe several of the many different structures which can exhibit the required controlled α. The high α transistor can be fabricated by conventional diffusion and/or alloy techniques. A controlled low-α structure may be built by reducing either the injection efficiency γ or the transport factor β to values near zero. A structure in which the low α is achieved by reduction of γ is shown in Fig. 2. MOLL and ROSS[4] and TANENBAUM and THOMAS[5] have shown that the injection efficiency of a transistor in which the diffusion length in the base is large compared to the base width (high β) and the diffusion length in the emitter is large compared to emitter width, and the surface recombination velocity at the emitter surface is infinite, is approximately

$$\gamma \simeq \frac{R_b}{R_e + R_b} \tag{6}$$

where R_b is the sheet resistance of the base layer and R_e is the sheet resistance of the emitter. Normally one makes $R_b \gg R_e$ to obtain high γ; to obtain a low γ one makes $R_e \gg R_b$. Typical

values for sheet resistances are indicated in the figure. High gain, both in turning on and turning off, is therefore achievable in the structure of Fig. 2.

The n–p–n is a conventional double-diffused

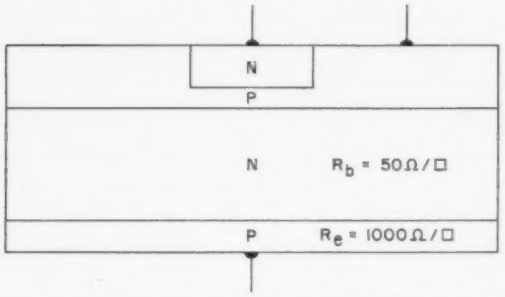


FIG. 2. Schematic representation of a p–n–p–n triode with high turn-off gain. The high α-transistor is the n–p–n made by conventional techniques. The low-α transistor is the p–n–p made with an emitter sheet resistance much higher than the base sheet resistance to provide a low injection efficiency.

silicon transistor. The p-type emitter is made by the diffusion of a high sheet resistance layer into the back of the n–p–n. Emission concentration does not alter the injection efficiency of this structure. By making this p-type emitter thin and highly doped, consistent of course with the requirement of high sheet resistance, conductivity modulation of the region, with a subsequent increase in injection efficiency, will not occur. Furthermore, the series resistance will be low. In this structure, lifetime is noncritical so long as it is high enough to provide diffusion lengths substantially greater than emitter and base widths.

A second structure somewhat similar to the first is shown in Fig. 3. In this case the lower emitter is limited in area as indicated and a shorting contact is placed across the emitter-base junction. The injection efficiency here is reduced by shunting the current around the emitter junction. If the sheet resistance of the n-layer is, for example, 1/20 the sheet resistance of the emitter layer, then approximately 95 per cent of the current into the n-layer will flow directly from the contact rather than by injection. For such a structure to be effective, however, the resistance of the n-layer must be large compared to the junction forward resistance.[6] The localized p^+ region can be used

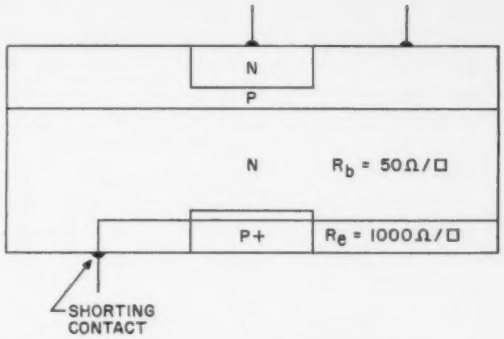


FIG. 3. Schematic representation of a p–n–p–n triode with high turn-off gain. The n–p–n is a conventional high α transistor. In the p–n–p the injection efficiency is reduced to a small value by means of the shorting contact across the emitter junction which causes most of the current to flow directly into the n-type base for the base and emitter sheet resistances indicated. The p^+-region insures a high local injection efficiency for that portion of the current which flows into the emitter from the shorting contact.

to insure that the local injection efficiency is near unity.

Fig. 4 shows a structure where the low α is achieved by lowering the transport factor β. This structure, though less amenable to calculation

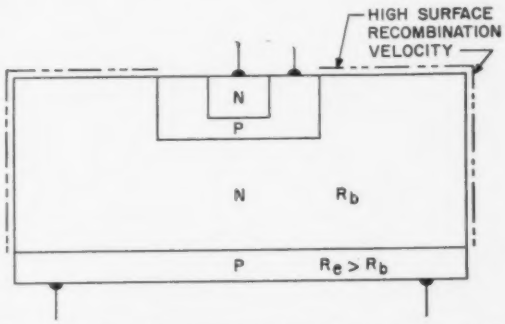


FIG. 4. Schematic representation of a p–n–p–n triode with high turn-off gain. The high-α transistor is a conventional n–p–n. The transport factor of the p–n–p is made low by forcing the emission to take place away from the collector junction and by providing high surface-recombination velocity surfaces.

than the others, can also be used to produce the desired effect. The β is reduced by making the collector junction much smaller than the emitter

and by providing high recombination velocity at the surface of the n-type base. Contacts are made to the edge of the emitter. Although the resistance of the n-type base layer will tend to shift the emission toward the center, the resistance of the emitter itself will tend to shift the emission away from the collector and back to the vicinity of the contacts. Thus, again the emitter sheet resistance should be higher than the base sheet resistance though the large ratios required above are not needed here.

EXPERIMENTAL RESULTS

Fig. 5 illustrates the turn-off gain of a typical unit made in the structure of that shown in Fig. 2. The turn-off gain is plotted as a function of the collector or load current in the figure. Note that

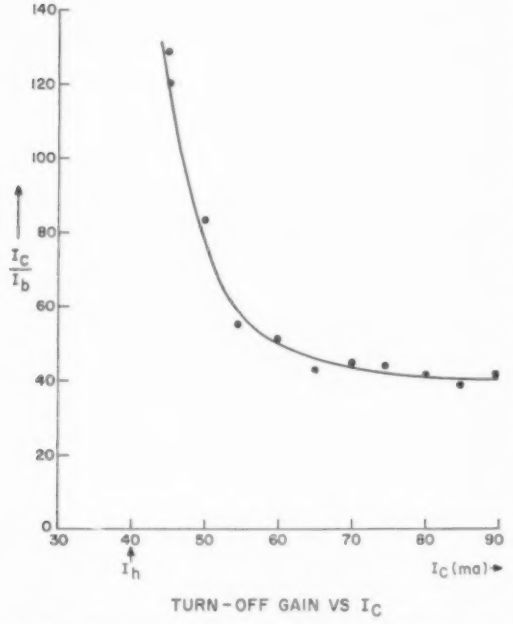


FIG. 5. Experimental data showing turn-off gain of a p–n–p–n triode of the structure of Fig. 2 as a function of collector current. The gain approaches infinity at the two-terminal holding current I_h, decreases with increasing collector current and saturates at a value of 40.

near the holding current the gain approaches infinity as it must when $\alpha_n + \alpha_p$ is just equal to unity. As the collector current increases the gain falls, as the n–p–n α increases, and then saturates at a level of about 40.

CONCLUSIONS

The basic requirement for the achievement of turn-off gain in p–n–p–n triodes is that the sum of the α's just exceed unity when the device is in the "on" or low impedance state. In order to achieve large turn-on as well as large turn-off gain the α of the component transistor to which the base lead is attached is made to approach unity while that of the other component transistor is made to approach zero.

Several structures in which large turn-off and turn-on gain can be achieved have been described. In each of these the high-α component transistor is made by conventional diffusion and/or alloy techniques. The low-α component transistor can be fabricated with a low-injection-efficiency emitter achieved by making the sheet resistance of the emitter much greater than that of the base. Alternatively, the transport factor may be made low. p–n–p–n switches made in this way are not driven heavily into saturation when turned on and as a result other effects will occur. First the storage portion of the turn-off time of these devices will be low because the excess stored charge is low. Rise time on the other hand will be increased. In addition, one may anticipate slightly larger "on" drop in high-voltage structures since most of the current through one base is carried by majority carriers.

REFERENCES

1. I. M. MACKINTOSH, *Proc. I.R.E.* **46,** 1229 (1958).
2. C. W. MUELLER and J. HILIBRAND, *IRE Trans. Electron Devices* **5,** 2 (1958).
3. J. J. EBERS, *Proc. I.R.E.* **40,** 1361 (1952).
4. J. L. MOLL and I. M. ROSS, *Proc. I.R.E.* **44,** 72 (1956).
5. M. TANENBAUM and D. E. THOMAS, *Bell Syst. Tech. J.* **35,** 1 (1956).
6. R. W. ALDRICH and N. HOLONYAK, JR., *J. Appl. Phys.* **30,** 1819 (1959).

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 123–126. Printed in Great Britain

A NEGATIVE-RESISTANCE PHOTOELECTRIC DIODE

W. RINDNER

Raytheon Research Division, Waltham, Massachusetts

(*Received* 21 *December* 1960)

Abstract—The fabrication and some of the properties of a silicon photoelectric diode displaying a negative-resistance region in its *I–V* characteristic are described. The sensitivity to light is particularly pronounced in the negative-resistance region and is determined by the amplitude and wavelength of the radiation applied.

Résumé—On décrit dans cet article la fabrication et quelques-unes des propriétés d'une diode photoélectrique au silicium qui manifestent une région à résistance négative dans la caractéristique *I–V*. La sensibilité à la lumière est particulièrement prononcée dans la région à résistance négative; elle est déterminée par l'amplitude et la longueur d'onde de la radiation appliquée.

Zusammenfassung—Die Herstellung und Eigenschaften einer photoelektrischen Siliziumdiode mit einem Bereich negativen Widerstandes in der *I–V*-Kennlinie werden beschrieben. Die Lichtempfindlichkeit ist in dem Bereich mit negativem Widerstand sehr ausgesprochen und wird durch die Amplitude und Wellenlänge der aufgewandten Strahlung bestimmt.

THE presence of a negative-resistance region in the *I–V* characteristic of a variety of two-terminal and multi-terminal semiconductor devices has in recent years attracted increasing interest. The usefulness of these devices centers around the particular transfer properties for electrical signals inherent in the presence of a negative resistance. The purpose of this paper is to report preliminary results obtained from a novel photoelectric diode. This device offers similar advantages in the transfer properties for light signals, in addition to displaying the desirable features of a conventional negative-resistance diode. The sensitivity to light is particularly pronounced in the negative-resistance region, whose slope and range are affected to a degree determined by the amplitude and wavelength of the radiation applied. These properties, as compared with those of conventional photoelectric devices, offer interesting possibilities for various applications.

The semiconductor material used is 1000 Ω-cm silicon, about 10 mils thick in the (111) direction, and with a surface area of about 30×30 mils. The chip is mounted on a heavily gold-plated JEDEC–30 header by suitable heating and formation of a Si–Au eutectic bond. A piece of aluminum wire 2 mils in diameter is attached to one of the header terminals by welding or thermocompression bonding. The sub-assembly is then boiled in Permachlor, etched in CP-4 for 15–20 sec, washed in de-ionized water and dried in a nitrogen atmosphere. The unit is then alternately connected to a condenser-discharge pulsing apparatus and an oscilloscope displaying its *I–V* characteristic. Mechanical contact between the Al wire and the semiconductor is established and the wire pulsed in. In most cases this results in a

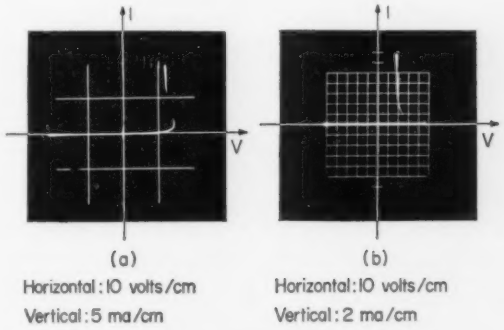


FIG. 1. Dark-state *I–V* characteristics.

typical characteristic as shown in Fig. 1(a). If instabilities are present they can often be removed by repeating the pulsing procedure. When satisfactory performance is obtained a metal ring is pressed over the rim of the header to protect the finished unit (Fig. 2) from handling.

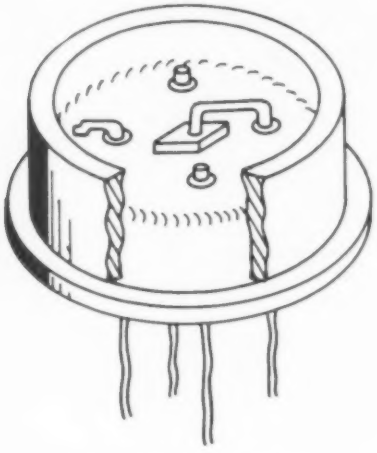


FIG. 2. The diode assembly.

Occasionally, turnover voltages V_T as high as 100 V were observed as in Fig. 1(b), or V_T values as low as 3 V. Various values of V_T have been obtained on the same chip by altering the pulsing procedure. Generally, it was found that V_T decreased as the contact size increased.

The spectral sensitivity at room temperature was determined within the range of about 0·4–1·4 μ. Care was taken to eliminate possible errors due to heating and drift. The effects of illumination were determined by observation of the I–V characteristic on a low-frequency curve tracer. Illumination has been found to alter the device conductance at all bias values, both positive (the Al contact positive) and negative and most significantly in the negative-resistance region. Fig. 3 shows the effect of illumination with white light of various intensities. Both V_T and the sustaining voltage V_S (minimum voltage) decrease with increasing intensity of white light. The negative-resistance region decreases and finally disappears as the illumination intensity increases. Spectrographic analysis shows that at room temperature this effect is most pronounced at about 1·02 μ. Fig. 4 shows the dependence of V_T on wavelength

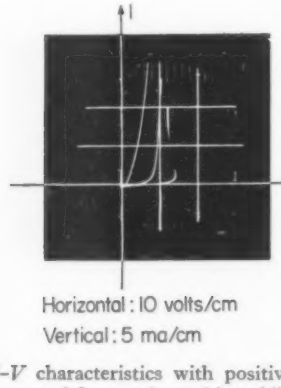


FIG. 3. I–V characteristics with positive bias for the dark state and for two intensities of illumination.

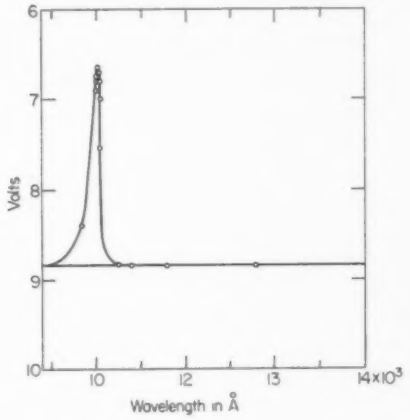


FIG. 4. Turnover voltage as a function of wavelength.

with constant light intensity. Fig. 5 shows the dependence of V_T on intensity at 1·02 μ.

The extent of the active device surface has not yet been determined and thus its absolute sensitivity is not known at present. In one case, with a unit biased close to its turnover voltage, a total radiant power of about 4 μW at a wavelength of 1·02 μ resulted in a change of 2·8 V and 0·65 mA across a load resistance. The radiant power input was derived assuming the whole device surface to be active. Occasionally very high sensitivities were observed while the I–V characteristic was displayed by a curve tracer. In these cases ordinary room illumination (about 10 ft from a 40-W

fluorescent lamp) resulted in decreases of V_T of the order of 10 V.

An estimate of the response time was obtained by exposing the unit to illumination from a tungsten filament in a suitable optical system. The light

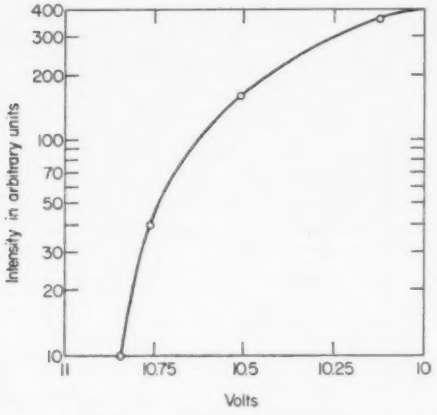


FIG. 5. Turnover voltage as a function of radiation intensity at a wavelength of about 10,200 Å.

beam was passed through a Kerr cell which was pulsed at a rate of 500 c/s with a decay time for the light pulse of about 2×10^{-7} sec. With the unit biased close to V_T the corresponding device response time, as observed on an oscilloscope across

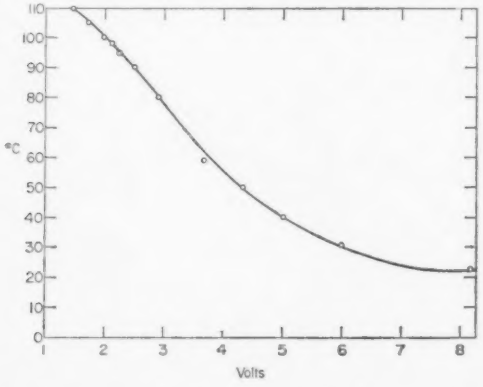


FIG. 6. Turnover voltage as a function of temperature.

the diode, was found to be about 2×10^{-6} sec. This probably constitutes a low estimate, since considerable capacity once external to the device was present when the measurement was made. No time lag in the response was observed with use of mechanical light-chopping techniques at frequencies of the order of 10 kc/s.

Determination of the temperature dependence of the device parameters between room temperature and about 110°C was performed by enclosing the unit in an open-air furnace. The I–V characteristic of the unit under positive bias was taken on a curve tracer at various temperatures. Increases of temperature were found to result in decreases of both V_T and V_S in a similar manner to increase of light intensity. At about 110°C the negative-resistance region disappears. Fig. 6 shows the relationship between V_T and temperature.

Admittance measurements were performed at 100 kc/s on a Boonton capacity bridge 74 C = S8 at a signal level of 100 mV r.m.s. In view of the relatively large signals applied, the results of these measurements must be interpreted with caution since some of the fine structure may have been masked. The measuring circuit is shown in Fig. 7.

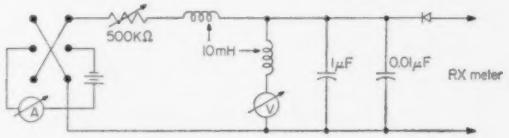


FIG. 7. The bridge measuring circuit.

The device admittance was measured with both bias polarities over a voltage range below the negative-resistance region. Considerable spread in the admittance values was observed on various units. However, within the bias range considered, a typical characteristic is represented by Fig. 8 for a unit with a V_T of 25 V. The values given include the package capacitance of about 1·1 pF.

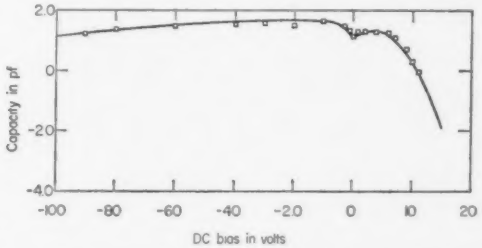


FIG. 8. Reactance as a function of bias at 100 kc/s.

The tendency to display inductive reactance under positive bias as the turnover voltage is approached, particularly at low frequencies, was also found under negative bias but at considerably higher voltages. Change of admittance towards inductive values was always accompanied by an increase of conductance.

Owing to the pulsing involved in the contact formation in conjunction with the high resistivity of the semiconductor used, the exact nature of the small junction is somewhat in doubt. It is evident that the aluminum contact determines the dark-state device characteristic to a large degree, as demonstrated by the strong dependence of the latter on pulsing procedure and consequent junction geometry. Current experiments suggest that injection from the large contact also plays a role in the device characteristics. It is conceivable that, under the small contact area, high fields are present which lead to current multiplication and avalanche breakdown. Such effects have in the past been associated with light sensitivity,[1] negative resistance[2, 3] and inductive reactance.[1, 4] It is hoped that work now in progress will clarify this point and permit the determination of an appropriate physical model.

In summary, the fabrication and some of the properties of a photoelectric diode displaying a differential negative resistance have been described. Potential applications of this device encompass, among others, light-controlled oscillators and switches actuated by light, as well as many functions of conventional photoelectric devices.

Acknowledgements—The author wishes to acknowledge helpful discussions with Drs. J. LAVINE and D. WARSCHAUER, the assistance received from R. NELSON and E. WANG in measurements, and from A. IANNINI in fabrication and measurements.

REFERENCES

1. B. SENITZKY and J. L. MOLL, *Phys. Rev.* **110,** 612 (1958).
2. J. B. GUNN, *Proc. Phys. Soc. Lond.* **69B,** 781 (1956).
3. A. F. GIBSON and J. R. MORGAN, *Solid-State Electron.* **1,** 54 (1960)
4. W. RINDNER, *Magnetic and Reactive Effects in a Negative-Resistance Diode.* To be published.

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 127–133. Printed in Great Britain

TRANSIENT PHOTOVOLTAIC RESPONSE OF DIFFUSED-JUNCTION SILICON PHOTODIODES*

T. E. HARTMAN

Texas Instruments Incorporated, Dallas, Texas

(*Received* 30 *March* 1961; *in revised form* 1 *May* 1961)

Abstract—The results of an experimental and theoretical study of the transient photovoltaic response of diffused-junction silicon photodiodes are presented. The analysis pertains to a thin *p*-type layer diffused in an *n*-type wafer illuminated on the *p*-type surface. According to an exact solution for the transient response of the photocurrent at the junction, the rise time is limited by minority-carrier storage. Expressions for the external photovoltage rise and fall times indicate that in general the rise time at room temperature with no external bias depends on the light intensity as well as the load resistance, while the fall time depends only on the load resistance. Previous attempts by others to show experimental agreement with theory have not been completely satisfactory. Some experimental difficulties have been overcome by using an extremely fast light chopper. Rise and fall time measurements made on a silicon photodiode (TI Type LS 222) are in very good agreement with analysis. Impedance measurements which are important in showing the effect of load resistance on rise time are also discussed.

Résumé—On présente les résultats d'une étude expérimentale et théorique de la réponse voltaïque transitoire des photodiodes au silicium aux jonctions diffusées. L'analyse se rapporte à une couche fine de type *p* diffusée dans une tranche de type *n* illuminée sur la surface du type *p*. D'après la solution exacte pour la réponse transitoire du courant photoélectrique à la jonction, la durée d'établissement est limitée par l'emmagasinage de porteurs minoritaires. Des équations pour la durée d'établissement et de chute de la tension photoélectrique externe indiquent qu'en général la durée d'établissement à température ambiante sans polarisation externe dépend de l'intensité lumineuse aussi bien que de la résistance de charge, tandis que la durée de chute dépend uniquement de la résistance de charge. Plusieurs essais précédents tentant de démontrer l'accord expérimental avec la théorie n'ont pas entièrement réussi. Certaines difficultés expérimentales ont été surmontées en utilisant un tranchant lumineux extrêmement rapide. Les mesures des durées d'établissement et de chute faites sur une photodiode au silicium (TI Type LS 222) sont en bon accord avec l'analyse. Les mesures d'impédance qui sont importantes pour demontrer l'effet de la résistance de charge sur la durée d'établissement sont aussi discutées.

Zusammenfassung—Der photoelektrische Sperrschichteffekt einer Siliziumphotodiode mit Diffusionsübergang wurde theoretisch und experimentell untersucht. Benutzt wurde eine dünne Schicht vom *p*-Typ, die in eine Schicht vom *n*-Typ eindiffundiert war, Beleuchtung erfolgte auf der *p*-Typ Oberfläche. Eine genaue Lösung für den Sperrschichteffekt des Photostroms am Übergang ergibt, dass die Anstiegszeit durch Anhäufung von Minoritatsträgern begrenzt wird. Ausdrücke für die Anstiegs- und Abfallszeit der äusseren Photospannung zeigen, dass die Anstiegszeit bei Zimmertemperatur. wenn keine Spannung angelegt wird, von der Lichtintensität und dem Belastungswiderstand abhängt, während die Abfallszeit nur vom Belastungswiderstand beeinflusst wird. Frühere Versuche anderer, eine experimentelle Übereinstimmung mit der Theorie zu erzielen, waren nicht sehr erfolgreich. Es gelang, einige der experimentellen Schwierigkeiten durch Verwendung eines sehr schnellen Lichtzerhackers zu überwinden. Die Messungen der Anstiegs und Abfallszeit bei einer Siliziumdiode (TI Typ LS 222) stimmen mit der Theorie gut überein. Impedanzmessungen, die den Effekt des Belastungswiderstandes auf die Anstiegszeit zeigen, werden ebenfalls erörtert.

* Presented at the Monterey meeting of the American Physical Society, March 1961.

1. INTRODUCTION

SEVERAL papers have appeared in the literature concerning the performance and physical mechanisms of semiconductor photovoltaic detectors.[1,2] The frequency response of grown,[1] diffused, and alloyed[3] p–n-junction photodiodes has been reported. The transient response has been qualitatively understood[4] and very recently the transport properties at the junction during transient irradiation were reported;[5,6] but a quantitative analysis of the observed photovoltage has been lacking. BECKER and FAN[7] observed the rise and fall of photovoltage under pulsed irradiation and concluded that the processes, while approximately exponential, were not simply the RC time constants. PRINCE and WOLF[4] made similar measurements but the long rise and fall times of the light pulse employed placed severe limitations on the experimental conditions and transient times which could be analyzed. However, they concluded that over the range of their observations, "the transient times depended mainly on capacitance–resistance effects".

The purpose of this paper is to present a quantitative analysis of the transient response times and to show experimental results over an extended range of conditions and transient times. A transient equivalent circuit is proposed in Section 2.2, analyzed in Section 2.3 and compared with experimental observations in Section 3.

2. GENERAL ANALYSIS

The diffused p–n-junction configuration considered in this paper is illuminated normal to the plane of the junction (i.e. parallel to the x-axis) on the p-type surface ($x = 0$). The structure consists of a very thin p-type region ($0 \leqslant x \leqslant W$) diffused into a thicker n-type region ($W \leqslant x \leqslant W+w$).

The following assumptions[1] have been made to simplify the analysis:

(1) The incident light is monochromatic and results in a quantum efficiency of unity.*

(2) There is an ohmic contact covering the entire surface at $x = W+w$ and the contact to the surface at $x = 0$ is of negligible area relative to the area irradiated by the light.

(3) Carrier trapping is negligible; i.e. the decay of excess carriers in the bulk can be characterized by the single exponential, $\exp(-t/\tau)$.

(4) The light intensity corresponds to low-level illumination, i.e. low injection level.

(5) The junction is abrupt, i.e. the transition region is small compared to W.

2.1 *Frequency response*†

An expression for the small-signal a.c. component of the photovoltage v_{ac} is given by SAWYER and REDIKER[3] for photovoltaic operation as

$$v_{ac} = \frac{qM(1-\psi)Q_i AF(\alpha, \omega)\exp(i\omega t)}{Y_{ac}+Y_L}$$

where q is the electronic charge, M the degree and ω the frequency of modulation of the incident light, ψ the surface reflection coefficient, Q_i the number of incident photons per cm^2 sec, A the illuminated surface area, $F(\alpha, \omega)$ a function independent of time, Y_L the admittance of the load, and Y_{ac} the internal admittance of the junction. For values of

$$\omega \ll \omega_0 = 2{\cdot}46D_n/W^2$$

where ω_0 is the cutoff frequency, the admittance Y_{ac} may be represented by a parallel combination of a resistance R_{ac} and a capacitance C_{ac} whose values for $sW/D_n \ll 1$ are

$$R_{ac} = (kT/qj_0A)\exp(-qV_{dc}/kT) \qquad (1)$$

and

$$C_{ac} = W(R_{ac}s)^{-1}$$

where D_n is the minority-carrier diffusion constant in the p-type region, s the surface recombination velocity, k Boltzmann's constant, T the absolute temperature and j_0 is the reverse saturation current density. The RC representation is valid for $\omega \leqslant s/W$.

The analysis results in the a.c. equivalent circuit for photovoltaic operation shown in Fig. 1, where j is the current density at the junction. C_0 is the capacitance and R_0 is the shunt resistance of the barrier in the dark. R' is the series resistance of the base structure, and Y_L is denoted by R_L in parallel with C'.

* By superposition, the results can be generalized to apply for any input spectrum.

† This section follows a development given in Ref. 3.

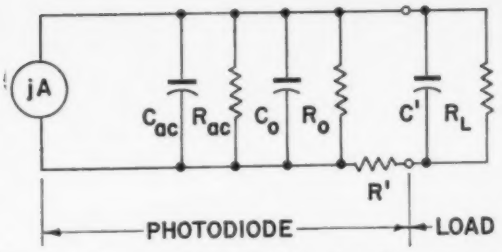


FIG. 1. A.C. equivalent circuit for photodiode operated in the photovoltaic mode.

2.2 *Transient response*

An expression for the time dependence of the current density at the junction when a step function pulse of light is applied to the photodiode has been given by GRINBERG and STROKAN[6] as

$$j(t) = 2xj_0W\exp(-\alpha W)\sum_{k=1}^{\infty}\left\{1-\exp\left[-\left(\Theta_k^2+\frac{W^2}{L_n^2}\right)\frac{D_nt}{W^2}\right]\right\}$$
$$\times\frac{\Theta_k^3\left[\Theta_k\sin\Theta_k+\exp(\alpha W)\left(\frac{s}{D_n}+\alpha\right)W-\frac{sW}{D_n}\cos\Theta_k\right]}{(\Theta_k^2+\alpha^2W^2)\left(\Theta_k^2+\frac{W^2}{L_n^2}\right)\left[\Theta_k^2+\frac{sW}{D_n}\left(1+\frac{sW}{D_n}\right)\right]\sin\Theta_k} \qquad (2)$$

where $L_n = (D_n\tau)^{1/2}$ is the minority-carrier diffusion length, α is the radiation absorption coefficient and Θ_k are the roots of the equation

$$\tan\Theta = -\frac{\Theta D_n}{sW}$$

GRINBERG and STROKAN, with the aid of equation (2) for $W = 0{\cdot}5L$, concluded that τ_a, the time required for the amplitude to reach 0·1 of the steady-state value from the instant the excitation is applied, is in the range

$$\frac{0{\cdot}01}{\omega_0} < \tau_a < \frac{0{\cdot}5}{\omega_0}$$

and that τ_b, the time required for the amplitude to change from 0·1 to 0·7 of the steady-state value, is in the interval

$$\frac{0{\cdot}5}{\omega_0} < \tau_b < \frac{1{\cdot}5}{\omega_0}$$

Since the photodiode signal approaches an exponential, it is convenient to consider τ_0, the rise time of the current at the junction,* where $\tau_0 = \tau_a+\tau_b$ is approximately equal to the time constant in the exponential $\exp(-t/\tau_0)$. Note that the greatest lower bound of τ_0 is the minority-carrier storage time.[8]

For the most part, τ_0 is small compared to the rise and fall times of the photovoltage observed across the load. When this condition is obeyed, the proposed transient equivalent circuit shown in Fig. 2 is applicable. Since the intrinsic junction processes come to quasi-equilibrium in the time τ_0, the parallel combination of R_{ac} and C_{ac} is assumed switched into the circuit during the rise of externally observed photovoltage, and out during the fall. As a first-order approximation, the rise and fall times of the internal current generator (i.e. τ_0) are neglected. A small effect due to τ_0 is observed in certain cases and will be noted subsequently. Equation (1) is the inverse of the slope of the I–V characteristic of the photodiode at $V = V_{dc}$, where for photovoltaic operation V_{dc}

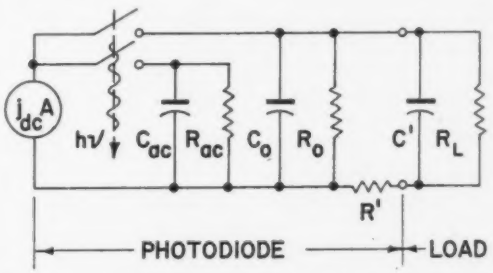


FIG. 2. Transient equivalent circuit for photodiode operated in the photovoltaic mode.

* The fall time of the current at the junction can be shown to be τ_0 also.

is produced by the generation of a non-equilibrium carrier concentration arising from radiation.[2]

For a given value of the load resistance (i.e. $R_L = 4{\cdot}52K$) the load line is shown on the I–V characteristics in Fig. 3. It is clear that R_{ac} increases as the light intensity decreases for R_L held constant.

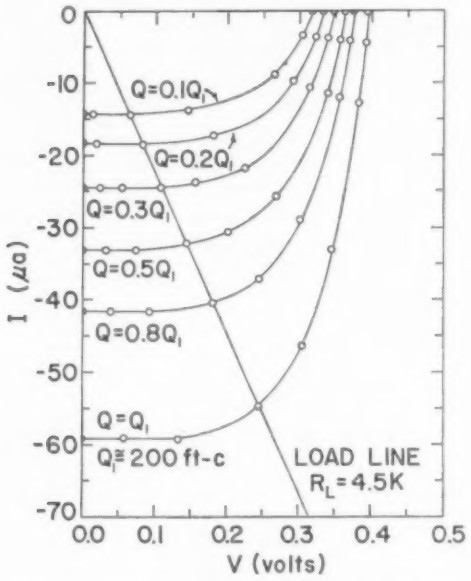


FIG. 3. Family of measured I–V characteristics for photovoltaic operation of the photodiode for various light intensities.

2.3 *Time constants*

The time dependence of the photovoltage across the load (i.e. R_L in parallel with C') is the same as that of the charge q on C'. The differential equation for q as a function of time is

$$a\ddot{q}+b\dot{q}+cq = RR_LC'j_{dc}A \tag{3}$$

where

$$a = RR_LR'CC',$$
$$b = RR_L(C+C')+R'(R_LC'+RC),$$
$$c = R+R_L+R',$$
$$R = R_{ac}R_0/(R_{ac}+R_0),$$
$$C = C_{ac}+C_0,$$

and j_{dc} is the steady-state junction photocurrent density. The general solution of equation (3) is

$$q = A_1\exp[-(\gamma-\beta)t]+A_2\exp[-(\gamma+\beta)t] + \frac{RR_LC'j_{dc}A}{R+R_L+R'}$$

where

$$\gamma = b/2a, \quad \beta = (b^2-4ac)^{1/2}/2a$$

The boundary conditions at $t = 0$ are

$$q = RR_Lj_{dc}A/(R+R_L+R'), \quad \dot{q} = 0 \tag{4}$$

The coefficients A_1 and A_2 may be evaluated using equation (4) and the absolute value of their ratio is

$$|A_1/A_2| = (\gamma+\beta)/(\gamma-\beta) \tag{5}$$

i.e. the absolute value of the ratio of the amplitudes of the responses is equal to the ratio of the associated time constants where

$$\tau_1 = (\gamma-\beta)^{-1}, \quad \tau_2 = (\gamma+\beta)^{-1} \tag{6}$$

Since $b^2 \gg 4ac$ for all R_L and R combinations involved then $\gamma+\beta \gg \gamma-\beta$ and from equation (5) $|A_1| \gg |A_2|$; i.e. the observed photovoltage exhibits essentially the single exponential with the longer time constant. Also, β may be expanded in a series and the leading terms of the expansion are given by

$$\beta \simeq \frac{1}{2a}\left(b-\frac{4ac}{2b}\right) \tag{7}$$

When equation (7) is substituted into equation (6) τ_1 becomes

$$\tau_1 \simeq \frac{RR_L(C+C')+R'(R_LC'+RC)}{R+R_L+R'} \tag{8}$$

The rise time of the photovoltage across the load, τ_r, is given approximately as

$$\tau_r \simeq \frac{R_{ac}R_L(C+C')}{R_{ac}+R_L}; \ (R_L > 1K) \tag{9}$$

by equation (8), since $R_{ac} \ll R_0$ except for small R_L (i.e. $R_L < 1K$) where R_{ac} is given by equation (1) and V_{dc} is the saturated (i.e. steady-state) photovoltage. The fall time is given approximately as

$$\tau_f \simeq \frac{R_0R_L(C_0+C')}{R_0+R_L}; \ (R_L > 1K) \tag{10}$$

by equation (8) since $R_0 \gg R' \ll R_L$, except for small R_L. The rise and fall times given by equations (8–10) are the times required for the amplitude to rise to $(1-e^{-1})$ and to fall to e^{-1} of the saturated value respectively.*

3. MEASUREMENTS

3.1 *A.C. impedance measurements*

An a.c. impedance comparison bridge was used to verify the equivalent circuit and to deduce the values of the parameters R_0 and R'. The impedance of the diode (TI type LS222) was compared with a resistance R_1 in parallel with a capacitance C_1. When the impedance of the diode is taken to be the equivalent circuit impedance, then at balance

$$R_1 = \frac{(R+R')^2+\omega^2C^2R^2R'^2}{(R+R')+\omega^2C^2R^2R'} \qquad (11)$$

where ω is the frequency of the a.c. signal generator. The results of measurements made on the diode in the dark are shown in Fig. 4 where the solid curve is a plot of equation (11), $R = R_0$ is $785K$, and R' is 46·3 Ω.

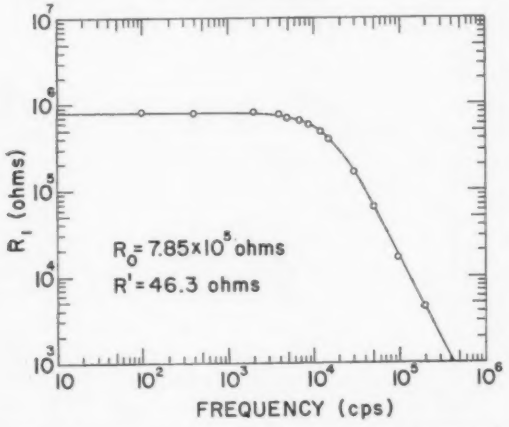


FIG. 4. Comparison of experimental and theoretical dispersion of R_1 where C' is 87 pF.

3.2 *Fall-time measurements*

A very fast light chopper developed by PEATTIE *et al.*[9] was used as a light source for all time measurements. The rise and fall times of the light pulse were approximately 10 mμ sec.† A typical oscilloscope trace for a fall time measurement is shown in Fig. 5. The results are shown in Fig. 6. The solid curve is a plot of equation (10) and the dashed curve is a plot of the more exact expression for τ_f given by equation (8) where $R = R_0$ and $C = C_0$. The range of fall times can be divided into three major regions:

(1) In the region where $R_L \gg R_0$ the fall time is independent of the load resistance, and is given by $\tau_f = R_0(C_0+C')$.

(2) In the region where $R_0 > R_L > \tau_0/C_0$ the fall time is given approximately by equation (10), and is strongly load dependent.

(3) In the region where $R' \gg R_L \ll \tau_0/C_0$, the fall time is independent of the load resistance, and is limited by τ_0 or $R'C_0$ whichever is larger.

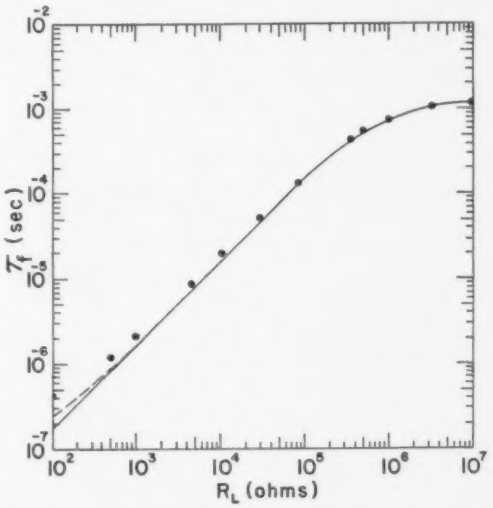


FIG. 6. Comparison of experimental and theoretical fall time dependence on load resistance.

The measured data in regions (1) and (2) indicated by the solid dots fit very well the theoretical curve generated by the present analysis. The

* Provided the diode response is allowed to saturate, the 10–90 per cent rise time and the 90–10 per cent fall time may be obtained by multiplying the appropriate τ by ln 9.

† The rise and fall times of the light pulse depend on the width of the surface to be illuminated. Since the entire sensitive area of the diode was illuminated (i.e. a width of 1·25 mm), the rise time is consequently long.

deviation of the measured values near region (3) is attributed to τ_0. τ_0 may be calculated using equation (2) and is about $0{\cdot}5$–$2{\cdot}0\times10^{-7}$ sec. The fall times were found to be independent of the light intensity over the entire range investigated (i.e. 25–500 ft-candles). This independence is predicted by equation (10), since R_0 does not depend on the light intensity, and is regarded as a further substantiation of the suggested model.

3.3 *Rise time measurements*

The "on-time" of the light pulse was adjusted to allow the photovoltage to saturate for rise time measurements. Care was exercised to maintain the rise time of the light less than 10 per cent of τ_r for all measurements; therefore, the accuracy of the rise time measurements was reduced to about 20 per cent. A typical oscilloscope trace for a rise time measurement is shown in Fig. 7. In addition to rise time measurements, data were taken for a static characteristic curve (i.e. the curve $Q = 0{\cdot}8Q_1$ in Fig. 3), and a.c. impedance measurements were taken concurrently. The results are shown in Fig. 8. The capacitance curve is a plot of $C+C'$ calculated from the impedance measurements. The curve of the junction resistance R is a plot of R_0 in parallel with R_{ac} where the latter is the inverse of the slope of the measured characteristic curve where it intersects the load line for R_L. The junction-resistance curve data were used to calculate the solid curve of R in parallel with R_L, and the open circles indicate the values obtained from impedance measurements. The curve of R in parallel with R_L was combined with the measured capacitance curve according to equation (9) to give the solid curve marked "rise time". The dashed curve is a plot of the more exact expression for τ_r given by equation (8). The range of rise times can be divided into three major regions:

(1) In the region where $R_L \gg R_{ac}$, the rise time is independent of the load resistance, and is given by $\tau_r = R_{ac}\,(C+C')$.

(2) In the region where $R_{ac} > R_L > \tau_0/C$, the rise time is given approximately by equation (9), and is strongly dependent on the load resistance.

(3) In the region where $R' \gg R_L \ll \tau_0/C$, the rise time is independent of the load resistance, and is limited by τ_0 or $R'C$ whichever is larger.

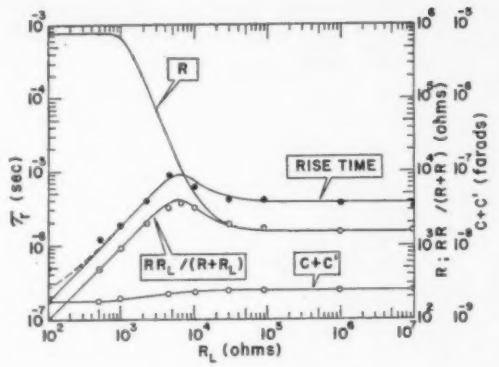


FIG. 8. Comparison of experimental and theoretical rise time dependence on load resistance where the light intensity is approximately 160 ft-candles.

The agreement in regions (1) and (2) between the calculated solid curve and the directly measured data indicated by the solid dots is within the limits of the estimated measurement error. The large deviation near region (3) again is attributed to τ_0. The effect of load resistance on rise time is clearly shown. As the magnitude of the load resistance is decreased, the junction resistance during illumination increases and R in parallel with R_L is a maximum at approximately the point of maximum power transfer to the load, i.e. where the junction resistance and the load resistance are equal. Since R is increasing more rapidly than R_L is decreasing, the maximum is displaced slightly. The maximum rise time occurs at approximately the same place, since the capacitance is changing only slightly. This result is also predicted by the present analysis, as shown by equation (9) when the change in capacitance is neglected. As the magnitude of the load resistance is reduced even further, short-circuit current operation is approached and R tends to R_0 as expected, since, from equation (1), R_{ac} becomes very large as V_{dc} approaches zero.

The rise times are dependent on the light intensity and are increased by decreased intensity. Excellent agreement is obtained between measured rise times and those calculated using equation (9), where R_{ac} is the inverse of the slope of the characteristic curves for different relative intensities at the points of intersection of these curves with the load line shown in Fig. 3.

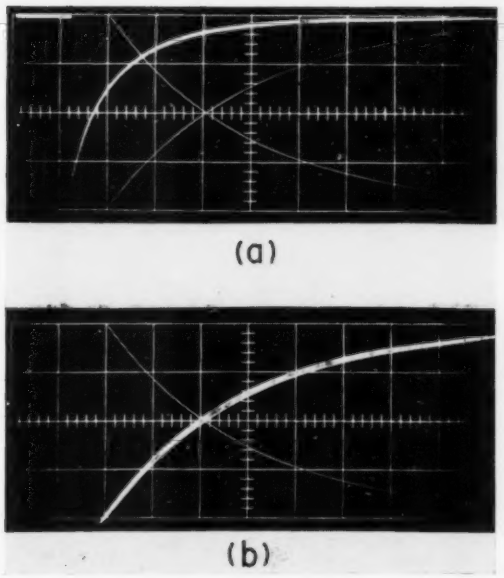


FIG. 5. Typical oscilloscope trace of the fall of photovoltage where R_L is 1 MΩ. (a) The whole response displayed with a long horizontal sweep rate. (b) The amplified fall response with the calibrated horizontal sweep rate adjusted to match the trace to a predetermined exponential engraved on the graticule. The response has been inverted by the electronics.

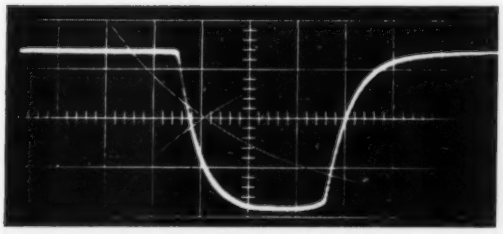


FIG. 7. Typical oscilloscope trace of the rise of photovoltage where R_L is 1 kΩ. The response has been inverted by the electronics.

[*facing p.* 132

4. CONCLUSIONS

The data presented show that the proposed transient equivalent circuit is valid particularly if the rise time of the current generator, τ_0, is taken into account. The rise time of the external photovoltage is dependent on the light intensity, while the fall time is not. Both are in excellent agreement with the analysis. The rise and fall times of the photovoltage across the load are for the most part simply the *RC* time constants of the equivalent circuit, except for small values of the load resistance where the transport properties of the junction become important.

REFERENCES

1. E. S. Rittner, *Photoconductivity Conference*, p. 215. Wiley, New York (1956).
2. R. L. Cummerow, *Phys. Rev.* **95,** 16 (1954).
3. D. E. Sawyer and R. H. Rediker, *Proc. I.R.E.* **46,** 1122 (1958).
4. M. B. Prince and M. Wolf, *J. Brit. Instn. Radio Engrs.* **18,** 583 (1958).
5. M. Green and I. N. Greenberg, *Solid-State Electron.* **1,** 85 (1960).
6. A. A. Grinberg and N. B. Strokan, *Fiz. Tverdogo Tela* **2,** 1536 (1960).
7. M. Becker and H. Y. Fan, *Phys. Rev.* **78,** 301 (1950).
8. J. L. Moll, *Proc. I.R.E.* **42,** 1773 (1954).
9. C. G. Peattie, W. J. Odom and E. D. Jackson, *Proc. Instn. Elect. Engrs.* **B106** (Suppl. 15), 303 (1959).

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 134–141. Printed in Great Britain

kMc/s PLANAR TRANSISTORS IN MICROWATT LOGIC CIRCUITRY

D. F. ALLISON, R. H. BEESON and R. M. SHULTZ

Fairchild Semiconductor Corporation, Palo Alto, California

(*Received* 13 *April* 1961)

Abstract—A new ultra-high-frequency transistor having a maximum frequency of oscillation of over 1 kMc/s and a gain-bandwidth product of about 800 Mc/s is described. It also has extremely low-leakage and usable d.c. current gain down to 1 μA and is seen to offer special advantages for application at extremely low current levels in logic circuitry. At the same time, owing mainly to the low collector capacitance of 0·6 pF (excluding the case), it is capable of operating at switching speeds of from 100 kc/s to 1 Mc/s at collector currents of 1–10 μA, respectively.

Diode–transistor logic, in general, and a new form using backward diodes, in particular, are seen to offer advantages in conservation of fan-in and fan-out at a given current level. This is especially important as the level is reduced. The latter form offers in addition a saving in number of components required.

D.C. conditions necessary to assure operation are considered, and it is seen that the circuits are feasible if the proper balance between diode and transistor voltages may be achieved. If not, then compensated circuits, using extra diodes either in series or as clamping elements, enable operation. The desirable features of a device which would satisfy the requirements are illustrated by the characteristic curve of an experimental silicon backward diode.

Performance is further demonstrated by a working four-stage shift register designed around a basic building block operating at 10 μA per transistor stage. The shift register operates satisfactorily at 200 kc/s from −40 to +100°C. It draws a total of about 360 μA. A computer using 10,000 such stages would draw a total current of about 100 mA at a power level of 300 mW.

Résumé—Un nouveau transistor d'hyperfréquence ayant une fréquence d'oscillation maximum de plus de 1000 Mc/s et un produit gain-largeur de bande d'environ 800 Mc/s est décrit. Cet élément a aussi un courant de fuite extrêmement faible et un gain de courant continu utilisable à moins d'un microampère et semble offrir certains avantages spéciaux dans les applications de circuits logiques à des niveaux de courants extrêmement bas. En même temps, en majeure partie due à la faible capacité du collecteur (environ 0,6 pF excluant l'enveloppe), il est possible de le faire opérer à des fréquences de commutation allant de 100 kc/s à 1 Mc/s à des courants de collecteur d'l à 10 μA respectivement.

En général, la logique des diodes-transistors et en particulier, une nouvelle forme employant des diodes à sens inverse offrent des avantages en conservant une ventilation régulière à un niveau de courant donné. Ceci est spécialement important si le niveau est réduit. Cette dernière forme offre une économie dans le nombre de pièces requises.

Les conditions de polarisation nécessaires pour assurer l'opération sont considérées et on remarque que les circuits sont réalisables si on observe un juste rapport entre les tensions des transistors et celles des diodes. Sinon des circuits de compensation employant d'autres diodes en série ou en éléments de blocage permettent l'operation. Les caractéristiques intéressantes d'un élément qui satisferait ces demandes sont illustrées par la courbe caractéristique d'une diode expérimentale à sans inverse au silicium.

Le rendement est aussi démontré par un enregistreur à déplacement à quatre étages dimensionné autour d'un bloc de base opérant à 10 μA par étage de transistor. L'enregistreur à déplacement opère d'une manière satisfaisante à 200 kc/s par sec entre −40 et 100°C. La charge totale est d'environ 360 μA. Une calculatrice utilisant 10 000 étages pareils tirerait un courant total d'environ 100 mA à un niveau de puissance de 300 mW.

Zusammenfassung—Ein neuer Höchstfrequenztransistor mit einer maximalen Schwingfrequenz von mehr als 1000 MHz und einem Produkt aus Verstärkung und Bandbreite von etwa 800 MHz wird

beschrieben. Er besitzt eine ausserordentlich niedrige Verlustrate und nützliche Gleichstromverstärkung bis zu 1 μA. Er eignet sich deshalb vorzüglich zur Anwendung bei äusserst niedrigen Stromniveaus in logischen Schaltkreisen. Glechzeitig ist er wegen der niedrigen Kollektorkapazität von 0,6 pF (ausschliesslich des Gehäuses) imstande, bei Kollektorströmen von 1 bis 10 μA mit Schaltgeschwindigkeiten von 100 kHz bis bzw. 1 MHz zu arbeiten.

Logische Schaltkreise von Diodentransistoren im allgemeinen und insbesondere eine neue Form, die Rückwärtsdioden benutzt, bieten den Vorteil der Aufrechterhaltung der Eingabe und Ausgabe bei gegebenem Stromniveau. Dies ist besonders bei Reduzierung des Niveaus wichtig. Bei der nuen Form ergibt sich ausserdem eine Einsparung an der Anzahl der Bauelemente.

Die zur Erzielung einer zuverlässigen Arbeitsweise erforderlichen Gleichstrombedingungen werden erörtert, und es ergibt sich, dass die Schaltkreise bei geeignetem Gleichgewicht zwischen Dioden und Transistorspannungen arbeitsfähig sind. Ist dies nicht zu erreichen, so müssen kompensierende Schaltkreise mit zusättzlichen Dioden, entweder in Reihenschaltung oder als aufgeklemmte Bauelemente, verwandet werden. Die erwünschten Eigenschaften eines Gerätes, das diesen Erfordernissen entspricht, werden durch die Kennliniex einer experimentellen Silizium-Rückwärtsdiode dargestellt.

Die Leistung wird weiterhin illustriert durch ein vierstufiges Register, dessen Grundlage ein mit 10 μA pro Transistorstufe arbeitender Bauelementblock bildet. Dieses arbeitet zwischen -40 und $+100$°C in zufriedenstellender Weise bei 200 kHz. Sein Strombedarf beträgt 360 μA. Eine Rechenanlage mit 10 000 derartigen Stufen würde bei einem Kraftniveau von 300 mW einen Gesamtstromverbrauch von etwa 100 mA haben.

INTRODUCTION

THIS paper will be concerned with the application of a new ultra-high-frequency silicon transistor to logic circuits operating at power levels in the microwatt region.

The device is extremely small and has a base width of about $\frac{1}{2}\mu$. It has a maximum frequency of oscillation in excess of 1 kMc/s, while the frequency of unity common-emitter-current gain is approximately 800 Mc/s. The collector capacitance, excluding that of the header, is about 0·6 pF.

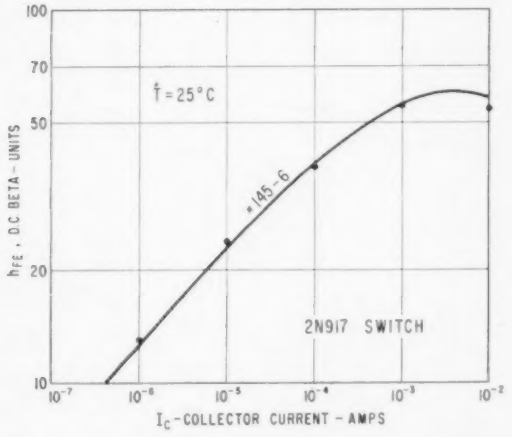


FIG. 1. Extended range of h_{FE} with very low collector current for UHF transistor.

By means of the planar process,[1] whereby the junction surface is located under oxide, excellent control of surface effects is obtained. The range of usable d.c. current gain, h_{FE}, is extended to below 1 μA as shown for a typical unit in Fig. 1, because of the reduction in surface recombination. Fig. 2 illustrates that the reduction in surface leakage causes a reduction in collector leakage current I_{CBO}, to values which are far below any operating level, even for microwatt logic.

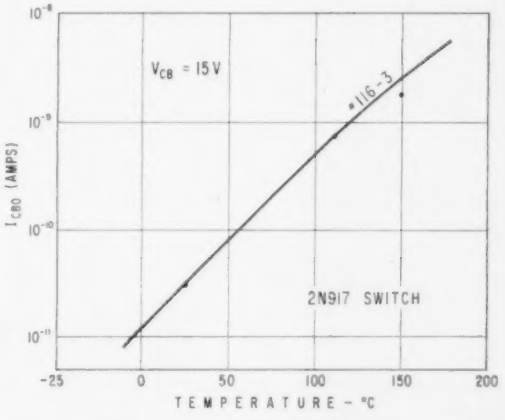


FIG. 2. Reverse collector current, I_{CBO}, is plotted against temperature to show that it is extremely low, even at high temperatures.

It will be shown that high-speed switching

circuits are possible at these low-current levels, primarily because of the low collector capacity and the high gain-bandwidth product. These will be capable of operating in the 100 kc/s to 1 Mc/s range at collector currents of from 1 to 10 μA, respectively.

NOTATION AND ABBREVIATIONS

Symbol	*Meaning*
V	voltage (V)
V_{BE}	base-emitter voltage (V)
V_{CB}	collector-base voltage (V)
V_{CE}	collector-emitter voltage (V)
V_{CC}	collector supply voltage (V)
I	current (A)
I_C	collector current (A)
I_B	base current (A)
I_E	emitter current (A)
I_{CBO}	collector-base leakage current (emitter open) (A)
T	temperature (°C)
P	power (W)
t_p	propagation time (from 50 per cent point on input to 50 per cent point on output) (sec)
f	frequency (c/s)
h_{FE}	d.c. current gain, or d.c. beta (dimensionless)
M	no. parallel inputs (fan-in) (dimensionless)
N	no. parallel outputs (fan-out) (dimensionless)

Subscript	*Meaning*
(sat)	in saturation
(on)	transistor in "on" state
(Th)	a threshold value
t	in tunnel direction
D	in diffusion direction

Abbreviation	*Meaning*
TRL	transistor-resistor logic
DCTL	direct-coupled transistor logic
BDTL	backward diode–transistor logic

BACKWARD DIODE LOGIC

The requirements of lightweight airborne or space equipment have been the main reason for the interest shown in microwatt-level circuits. Here the desire for low-current drain or low power is augmented by the desirability of simplicity in the logic so as to keep the number of components, and therefore the weight, to a minimum. Compared to other forms of logic, diode–transistor[2,3] logic is most desirable from this point of view, because of what might be termed conservation of fan-in or fan-out at a given current level. That is to say that in other forms of logic, such as TRL or DCTL, excess drive current must be supplied to a given base node to make up for signal "leakage" away from the node. In TRL, the signal leakage is back to other fan-in stages which may be on, and in DCTL the leakage is to other bases in parallel which "rob" the drive current available to a given base. Thus it may be said that for a given current level per stage, diode–transistor logic will give the highest fan-in and fan-out, since there are practically no problems of leakage of available base drive current. If the term "efficiency" may be said to apply to this conservation of fan-in and fan-out at a given current level, then diode–transistor logic may be said to be the most efficient form of transistor logic presently available.

A further simplification of diode logic is offered by the use of backward diodes,[4] such as is shown in the "NOT AND" logic circuit of Fig. 3. Here

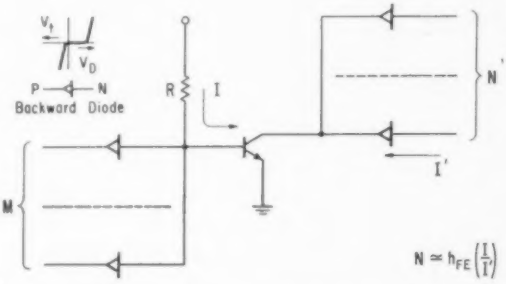


FIG. 3. Backward diodes used in a "NOT AND" logic circuit.

it is seen that there is no need for the voltage-translating or compensating devices necessary for turn-off in conventional diode–transistor logic. These usually consist of a diode, or else a resistor with a parallel capacitor, in series with the base with the addition of a negative supply to compensate for the drop through the diode. These components can be eliminated if backward diodes, which have a much lower drop, are used.

In Fig. 3 and henceforth, the terms V_t, or voltage in the tunneling direction, and V_D, or voltage in the diffusion direction, are used to avoid confusion. The arrow points in the "easy flow" direction.

The "NOR" configuration of Fig. 4 is still another possibility, but it may exhibit difficulties

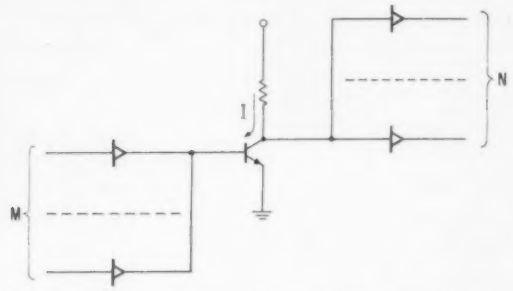


FIG. 4. Backward diodes used in a "NOR" logic circuit.

of base-current robbing, as previously described in connection with DCTL. However, this can be kept to a minimum as long as the collectors are not tied in parallel, which constitutes the worst case condition for this effect in DCTL.

DESIGN CONSIDERATIONS

The d.c. conditions under which these microwatt circuits will operate is of primary consideration. The formula for fan-out given in Fig. 3 is derived in Appendix A. It shows that fan-out is dependent upon d.c. beta and a "current source factor". The current source factor is the ratio of

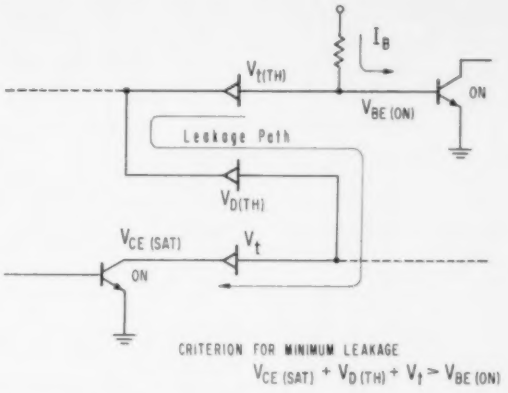


FIG. 5. Leakage paths in BDTL (backward diode transistor logic) circuitry.

available turn-on base current to collector current and is largely dependent on supply voltage. Very low supply voltages can be used when d.c. beta is high; however, it will be found that there is a loss of speed below a certain critical voltage. This results from the fact that as voltage is reduced there is a loss in turn-on drive current which causes a deterioration in rise time. For example, the circuits to be described later exhibit a loss in speed below 3 V, though it would be possible to operate them down to about 1 V while retaining a fan-out of 3 or 4, if we assume a minimum beta of of about 10. This, however, assumes no leakage through a path such as that of Fig. 5. The criterion necessary to keep this leakage to a minimum,

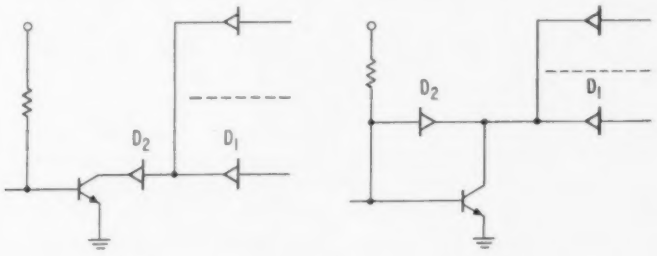


FIG. 6(a). Compensating circuit that increases $V_{CE(sat)}$ by adding another V_t term. (b) Compensating circuit that increases $V_{CE(sat)}$ by clamping the collector.

shown in the figure, is developed in Appendix B. From this formula it appears that high values of $V_{CE(sat)}$, V_D and V_t are desirable. This conflicts with the need to keep $V_{CE(sat)}$ and V_t low for turn-off of the following stage, however. A compromise must therefore be made.

Since the changes in $V_{D(Th)}$ and $V_{BE(on)}$ will be in the same direction with temperature, and since V_t changes very little, the change of $V_{CE(sat)}$ with temperature will have the greatest effect on this criterion.

If the above conditions cannot be satisfied, then it would be possible to use two diodes in series in place of one. Even this expedient would offer advantages over ordinary diode logic, but, alternatively, a saving in diodes may be had by using circuits such as those of Fig. 6.

The criterion shown in Fig. 5 then becomes that shown in Fig. 6(a) for this configuration. It can be seen that this effectively increases $V_{CE(sat)}$ by adding another V_t term. This formula will be strictly true only if the two diodes are of the same type. It may be advantageous in some cases to increase the size of the drop in the tunneling direction of diode no. 2, and thus greatly increase the effective $V_{CE(sat)}$.

The configuration of Fig. 6(b) effectively raises $V_{CE(sat)}$ by clamping the collector and thus not allowing it to go into saturation. It is shown in Appendix C that in this case the condition of Fig. 5 will be satisfied by the first criterion of Fig. 6(b).

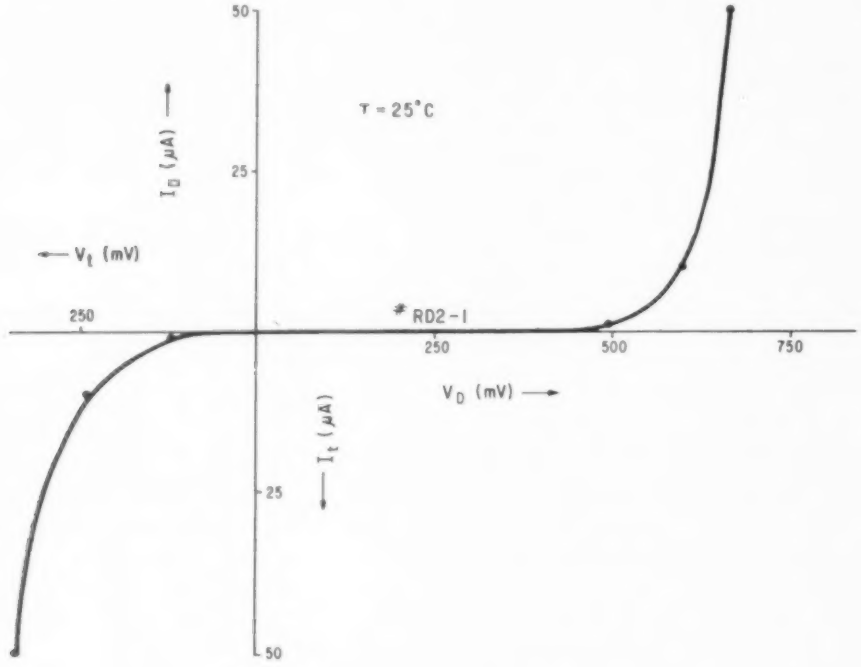


FIG. 7. Characteristics of an experimental silicon backward diode.

This puts the burden of satisfaction of this requirement entirely on the diodes, and should be quite easy to meet. A second restriction is put on the diodes, however; this is necessary to assure turn-off of the following stage. In other words, it must be assured that the resulting effective $V_{CE(sat)}$ is low enough to assure turn-off of the next stage. This is also shown in Appendix C.

To illustrate the characteristics desired in a backward diode for operation at microampere levels, the characteristic curve of an experimental silicon backward diode is shown in Fig. 7. These diodes do not display any noticeable negative-

resistance region, as in a tunnel diode. This would actually be detrimental to proper operation of the circuits, since there would be a high value of leakage due to tunneling current in the diffusion direction. The diodes have been made over a wide range of tunnel voltages.

RESULTS OF SWITCHING TESTS

Fig. 8 shows average propagation time per stage through ten single stages vs. current for the BDTL (backward diode transistor logic) circuit using the

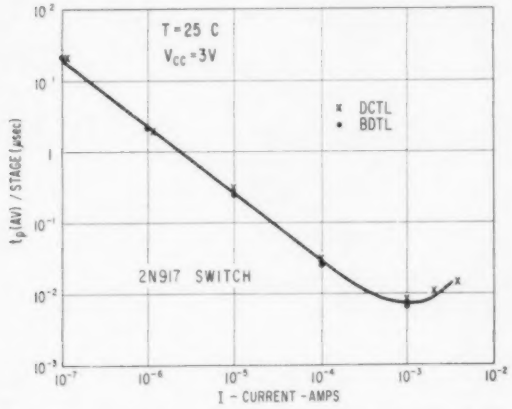


FIG. 8. Average propagation delay time per stage, through ten stages, vs. current.

kMc/s planar transistor and experimental backward diodes, as compared with DCTL. These are found to be nearly identical. Above about 1 mA, speed is limited by storage, while below about 1 mA, it is limited by fall time. This is due to the long charge-up time of the collector capacitance and associated capacitances to ground through the resistor. Low capacity is, therefore, essential for high-speed operation at low currents.[5]

An important point to realize regarding the backward-diode logic, as opposed to DCTL, is that this curve represents the worst propagation time at low currents, since increased fan-out *decreases* propagation time. The reason for this is that the current available to charge the circuit capacitances is increased with fan-out. The further possibility that higher-level stages would give a decrease in propagation time also suggests itself, and this has been found to be the case. In this way, a higher-level inverter can be used in a manner similar to the way in which a emitter follower is used, to give increased fan-out with little loss in speed, or to step up the current level for going from low- to high-level operation.

Simulated fan-in had very little effect on speed.

Several flip-flops were built and gave the following results:

Current Level (μA)	Switching Rate (Mc/s)
100	5
10	1
1	0·15

The decrease in switching rate with current is quite noticeable, as should be expected.[6,7]

10 μA SHIFT REGISTER

In order to further demonstrate the feasibility of these circuits at low levels, a four-stage shift register of the type shown in Fig. 9 was built. This circuit was developed from a basic 10 μA backward-diode transistor logic block, from which most of the diodes were eliminated. The resulting circuit is a combination of DCTL and backward-diode circuitry; the two are compatible as long as no two bases are connected together.

Operation is such that, if a given stage contains a "one" when a reset pulse arrives, it will trigger a one shot which will override the reset pulse on the next stage and thus set it to "one".

This circuit has been found to operate satisfactorily, circulating all possible combinations of ones and zeros, at 200 kc/s over a temperature range of from $-40°C$ to $+100°C$.

Some difficulty was encountered with feed-through of fast signals, since the collector capacitance, diode capacitance, and resistor form a high-pass filter with increasingly lower cut-off frequency as the resistor is increased. However, if the rise time of the driving circuits is kept low, this does not present a problem. Also, low capacitance is beneficial in this regard.

There was also considerable trouble with radiated noise pick-up, but proper shielding entirely eliminated this problem. The final enclosed circuit is practically impervious to external noise.

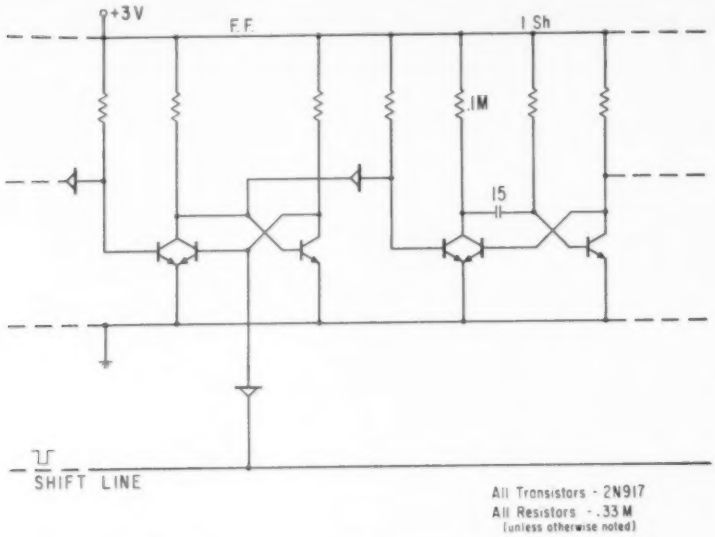


FIG. 9. One stage of a low-level four-stage shift-register design.

The four-stage shift register draws about 360 μA of current. A computer using 10,000 transistor building blocks of this type would draw a total current of about 100 mA, at a power level of 300 mW.

Acknowledgements—The authors are indebted to many persons for their help in gathering this material, and would like particularly to thank C. T. SAH for the backward diodes and B. J. BARRANGER for his work on the circuits.

APPENDIX A

From Fig. 3, considering only the collector node, we must require

$$I h_{FE} \geqslant N I'$$

$$N \leqslant h_{FE}\left(\frac{I}{I'}\right) \tag{A.1}$$

or

$$N \leqslant h_{FE}\left(\frac{V_{cc}-V_{BE(on)}}{V_{cc}-(V_{CE(sat)}+V_t)}\right) \tag{A.2}$$

APPENDIX B

Tracing the leakage path of Fig. 5, we have

$$V_{CE(sat)}+V_t+V_{D(Th)}+V_{t(Th)} > V_{BE(on)} \tag{B.1}$$

where $V_{D(Th)}$ and $V_{t(Th)}$ are defined as threshold values, specified at a given allowable leakage current. If we assume the worst case, however, $V_{t(Th)}$ might be very small, so that we have approximately

$$V_{CE(sat)}+V_t+V_{D(Th)} > V_{BE(on)} \tag{B.2}$$

APPENDIX C

In the case of Fig. 6(b), we have

$$V_{CE(sat)} = V_{BE(on)}-V_{t_2}$$

Now substituting this into equation (B.2)

$$V_{BE(on)}-V_{t_2}+V_{D_1(Th)}+V_{t_1} > V_{BE(on)}$$

$$V_{D_1(Th)}+V_{t_1} > V_{t_2} \tag{C.1}$$

Therefore, it is adequate to specify

$$V_{D_1(Th)} > V_{t_2} \tag{C.2}$$

For turn off, however, we must specify

$$V_{CE(sat)}+V_{t_1} < V_{BE(Th)}$$

substituting for $V_{CE(sat)}$;

$$V_{BE(on)}-V_{t_2}+V_{t_1} < V_{BE(Th)}$$

$$V_{t_2} > V_{t_1}+[V_{BE(on)}-V_{BE(Th)}] \tag{C.3}$$

therefore we must require at least that

$$V_{t_2} > V_{t_1} \tag{C.4}$$

REFERENCES

1. J. A. Hoerni, Presented at Electronics Development Meeting, Washington, D.C. (1960).
2. W. B. Cagle and W. H. Chen, *I.R.E. Wescon Convention Record*, Pt. 2, p. 3 (1957).
3. D. P. Masher, *I.R.E. Trans. Electron. Comp.* **EC-9,** 15 (1960).
4. I. A. Lesk, N. Holonyak, Jr., U. S. Davidsohn and M. W. Aarons, *I.R.E. Wescon Convention Record*, Pt. 3, p. 9 (1959).
5. C. D. Simmons, *Solid State Journal* **1,** 31 (1960).
6. H. Raillard and J. J. Suran, *Proc. I.R.E.* **47,** 96 (1959).
7. J. J. Suran, *Proc. I.R.E.* **49,** 420 (1961).

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 142–154. Printed in Great Britain

CONSTITUTIONAL SUPERCOOLING DURING CRYSTAL GROWTH FROM STIRRED MELTS—II

EXPERIMENTAL: GALLIUM-DOPED GERMANIUM

W. BARDSLEY, J. M. CALLAN, H. A. CHEDZEY and D. T. J. HURLE

Ministry of Aviation, Royal Radar Establishment, Great Malvern, Worcestershire, England

(*Received* 16 *March* 1961; *in revised form* 16 *May* 1961)

Abstract—The theory developed in a previous paper by HURLE, predicting the onset of constitutional supercooling during growth from a stirred melt, is verified experimentally for the case of the growth of ⟨110⟩ oriented single crystals of germanium from a gallium-doped melt. Contrary to the observations of BOLLING *et al.* it is found that the critical gradient of supercooling necessary to initiate cell formation is small or zero. The {110} cellular interface is briefly described.

Some published work relating to certain imperfections in crystals of germanium grown from doped melts is reviewed and it is shown that the nature of the imperfections is both qualitatively and quantitatively consistent with their being due to the existence of constitutional supercooling during the growth of the crystals.

Résumé—La théorie développée par HURLE dans un précédent article qui prédisait le début de la surfusion de constitution durant le développement en partant de la fonte remuée est vérifiée expérimentalement pour le cas du développement de cristaux orientés simples ⟨110⟩ de germanium en partant de la fonte dopée au gallium. Contrairement aux observations de BOLLING *et al.*, on a trouvé que la pente critique de la surfusion nécessaire pour initier la formation de cellule est très faible ou zéro. L'interface cellulaire {110} est brièvement décrite.

On revoit aussi certains travaux publiés qui sont relatifs à certaines imperfections dans les cristaux de germanium developpés des fontes dopées et on démontre que la nature de ces imperfections est qualitativement et quantitativement consistante à ce qu'elle est due à l'existence de surfusion de constitution durant de développement des cristaux.

Zusammenfassung—Die in einem früheren Artikel von HURLE entwickelte Theorie, nach der beim Kristallwachstum aus einer gerührten Schmelze Unterkühlung stattfindet wurde im Fall der Züchtung von längs der ⟨110⟩-Achse orientierten Germaniumeinkristallen aus einer mit Gallium dotierten Schmelze bestätigt. Im Gegensatz zu den Beobachtungen von BOLLING und Mitarbeitern ergab sich, dass der kritische Unterkühlungsgradient, der zum Einsetzen der Keimbildung erforderlich ist, klein oder Null ist. Die {110} Grenzfläche der Keimkristalle wird kurz beschrieben.

Existierende Veröffentlichungen über gewisse Fehlstellen in aus dotierten Schmelzen gezogenen Germaniumkristallen werden besprochen, und es wird gezeigt, dass die Natur der Fehlstellen ihrer qualitativen und quantitativen Beschaffenheit nach auf das Auftreten von Unterkühlung während des Kristallwachstums zurückgeführt werden kann.

1. INTRODUCTION

IN A previous paper[1] the theory of constitutional supercooling was applied to the case of crystal growth from a stirred melt. In the present paper the results of some experiments carried out to test the theory are given together with a brief description of the cellular structure.

The following expression was obtained,[1] for $(dS/dx)_{x=0}$, the gradient of constitutional supercooling in the melt at the interface between a rotating crystal and its melt,

$$\left(\frac{dS}{dx}\right)_{x=0} = -\frac{mfC_s(1-k)}{Dk} - \frac{K_sG_s - L\rho f}{K_L} \quad (1)$$

where

m = slope of the liquidus line from the phase diagram for the binary alloy
f = growth rate
D = diffusion coefficient for the solute in the molten solvent
k = distribution coefficient
C_s = concentration of solute in the solid at the interface
K_s, K_L = thermal conductivities of the solid and liquid respectively at the equilibrium temperature
L = latent heat of fusion of the solvent
ρ = density of the solvent
S = supercooling

Equation (1) can be expressed in terms of C_L, the concentration of solute in the bulk melt, by the equation due to BURTON *et al.*,[2]

$$C_s = \frac{k}{k+(1-k)\exp(-\Delta)} C_L \qquad (2)$$

where $\Delta = f\delta/D$, $\delta = 1{\cdot}6\, D^{1/3}\nu^{1/6}\omega^{-1/2}$, is the thickness of the diffusion-dominated boundary layer. ω is the angular velocity of the crystal and ν the kinematic viscosity of the melt.

The first and second terms on the right-hand side of equation (1) represent the slope, in the melt at the interface, of the equilibrium liquidus temperature distribution $(dT_E/dx)_{x=0}$ and the actual temperature distribution (G_L) respectively.

The onset of constitutional supercooling during the growth of a metal single crystal from an unstirred melt is known to produce a change in the morphology of the solid–liquid interface from a featureless to a cellular form.[3] Intermediate between the cellular and featureless forms, the interface becomes pocked.[4,5] The cells, which can be seen on the solid growth surface from which the melt has been rapidly decanted, are perpetuated through the crystal and hence the crystal consists of a number of parallel elements of prismatic form. The boundaries between the cells are enriched or depleted in solute according as $k \lesssim 1$ respectively.[3] Markings on the free surface of the crystal called corrugations can be seen where pairs of elements intersect the free surface. The published work relating to the effects of constitutional supercooling during growth from the melt has been reviewed by ELBAUM.[6]

The effects of constitutional supercooling on the solid–liquid interface morphologies of the semiconducting materials have not been extensively studied, the only work being that of BOLLING *et al.*[7] These authors investigated qualitatively the effects of constitutional supercooling on the growth, in a horizontal boat, of single crystals of germanium doped with gallium.

To the authors' knowledge, the effects of constitutional supercooling during crystal growth from a stirred melt has not been investigated previously. The advantage of stirring is that is imposes an impurity distribution in the melt which can be calculated and which, if the stirring is adequate, is not affected appreciably by any convective mixing in the melt.

The effects of constitutional supercooling on the growth of semiconducting materials obviously merit further study particularly in view of the present-day requirements for degenerately doped material for ESAKI diodes.[8] Accordingly, we have studied cell formation in germanium single crystals doped with gallium which were grown, by the Czochralski technique, from a stirred melt.

2. EXPERIMENTAL TECHNIQUES

A Czochralski crystal puller, developed in this laboratory to produce high-quality single crystals for research purposes, was used to grow oriented single crystals having a wide range of gallium concentrations. Each crystal was grown from a 100 g melt in an ambient atmosphere of dry argon and a variety of pull rates, ranging from 8×10^{-5} to $1{\cdot}75\times10^{-3}$ in/sec with crystal rotation rates of 10, 70 and 150 rev/min, were employed. The work reported here is concerned mainly with crystals grown on the $\langle 110\rangle$ axis, but crystals have also been grown on other axes. A $\langle 110\rangle$ growth axis was chosen so that the development of cellular structure and the resultant dislocation configurations could be traced down the crystal with the aid of suitable etching techniques.

With the high concentrations of gallium used, it was found necessary to remove all traces of water vapour and/or oxygen from the working chamber, otherwise reaction with the hot gallium coated the cooler parts of the chamber with a brownish powder. Therefore the apparatus was evacuated through a liquid-oxygen trap to a pressure of 0·03 mm Hg for a period of at least

30 min to drive off moisture absorbed in the graphite. Argon, which had been passed through a liquid-oxygen trap, was then flushed through the system for at least 15 min before the charge was melted and crystal growth commenced. The total weight of germanium/gallium was determined before and after crystal growth to check that no significant loss of gallium had occurred.

The germanium was zone-refined material having an extrinsic carrier concentration of the order of 5×10^{12} cm^{-3} and was etched in CP-4 etchant and carefully dried immediately before use. The gallium was quoted by the supplier (L. Light & Co.) to be 99·999 per cent pure and was further cleaned and purified by crystallizing it under dilute hydrochloric acid.

Every effort was made to keep the crystals all of the same size and shape. The crystals were grown out smoothly from the seed crystal to a constant diameter and, throughout the growth, rapid or large-scale changes in power to the melt were avoided.

For each growth rate the gallium concentration was adjusted so that corrugations first appeared on the crystal surface when between ~ 40 and ~ 60 per cent of the crystal had been grown. (Fig. 1 is a photograph of part of a $\langle 110 \rangle$ oriented crystal showing corrugations.) To establish that the corrugations were a visible manifestation of a cellular interface, the following experiment was performed. A growing crystal was rapidly withdrawn from the melt approximately 3 mm after corrugations appeared on the surface and it was noted that the interface was cellular. The experiment was repeated under identical conditions, only this time the crystal was rapidly withdrawn from the melt at a point just before corrugations would appear. The interface was then found to be smooth with no evidence of cells. Fig. 2 is a photograph of a cellular interface. It was therefore concluded that surface corrugations marked the onset of cellular growth.

All the crystals were cut perpendicular to the growth axis, with a diamond wheel, at the point at which corrugations appeared on the free surface, and the fraction of the total charge which had been crystallized when corrugations appeared was determined by weighing. The top halves of the crystals (the precellular regions) were then sectioned along one of the {111} planes parallel to the axis of growth, etched and examined to check that the crystals were single. The etchant used was one developed by CUMMEROW and CHERRY[9] [HF–HNO_3 (1:3) at 70°C for about 2 sec] specifically for showing dislocation etch pits in degenerately doped germanium. The optimum etching time was found to be ~ 15 sec (considerably longer than that stated by CUMMEROW and CHERRY[9]). Alternatively, the etchant was used cold with an etching time of ~ 4 min.

The concentration of gallium in the solid at the point at which corrugations began to form (C_{s_g}) was evaluated using the normal freeze equation[10]

$$C_{s_g} = k_{eff} C_{L_0} (1-g)^{k_{eff}-1} \tag{3}$$

where C_{L_0} = initial concentration of gallium in the melt and g = fraction of melt solidified. The effective distribution coefficient (k_{eff}) was calculated using the BURTON *et al.* formula[2]

$$k_{eff} = \frac{k}{k+(1-k)\exp(-\Delta)} \tag{4}$$

The true growth rate (f) at the point at which corrugations first formed was computed from the pull rate (f^1) and the rate of fall of liquid level in the crucible. It can easily be shown that

$$f = \left(\frac{r_m^2}{r_m^2 - r_c^2}\right) f^1 \tag{5}$$

where r_m = radius of the crucible in the plane of the solid–liquid interface and r_c = radius of the crystal.

The radius of the crystal was measured directly and the radius of the melt surface at the point at which corrugations formed was calculated from a knowledge of the profile of the crucible.

3. RESULTS

The relevant experimental data for the crystals grown is shown in Table 1.

Referring to equation (1) and making the assumption that a fixed critical value of $(dS/dx)_{x=0}$ is necessary to initiate cell formation, it can be seen that a graph of $\ln C_{s_g}$ vs. $\ln f$ should be linear at slow growth rates with a slope of -1 but that C_{s_g} should fall off more rapidly at fast rates of growth when the term $L\rho f$ becomes significant compared

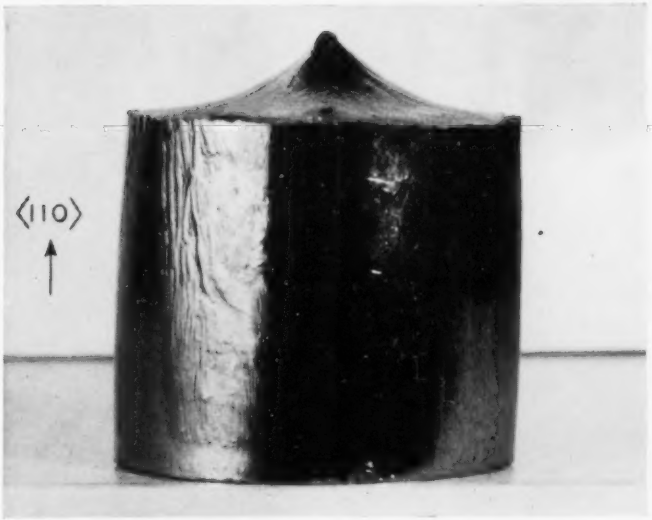


FIG. 1. Bottom half of a ⟨110⟩ oriented gallium-doped germanium crystal grown under conditions of constitutional supercooling showing corrugations. Magnification ×2·8.

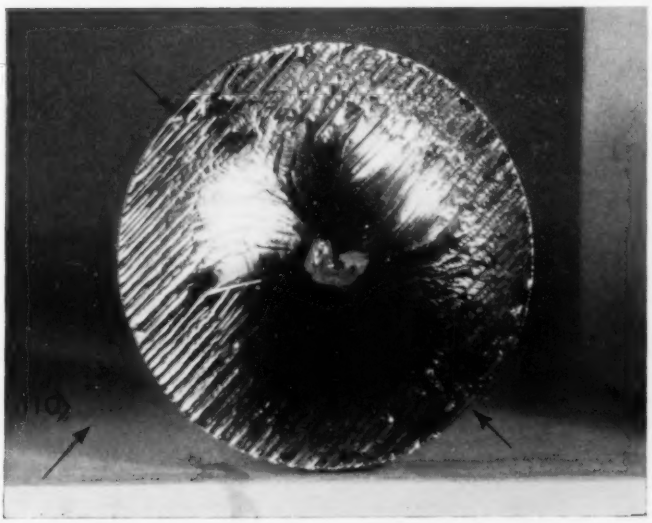


FIG. 2. Growth surface of ⟨110⟩ oriented gallium-doped germanium crystal just after corrugations had appeared, showing cellular interface. Arrows indicate positions of small {111} facets. Magnification ×2·8.

[*facing p.* 144

Table 1*

Crystal no.	Growth rate f (10^{-3} cm sec^{-1})	Rotation rate ω (rev/min)	Initial conc. of Ga in melt, C_{L_0} (at. per cent)	Fraction solidified when cells formed g	Conc. in liquid when cells formed C_{Lg} (at. per cent)	Conc. in solid when cells formed C_{sg} (at. per cent)	p^1C_s (°C)	p^1C_sf (10^{-3}°C cm sec^{-1})
HC 92	1·96	70	0·5970	0·593	1·34	0·130	5·19	10·17
HC 95	1·09	70	1·198	0·601	2·78	0·234	10·08	10·99
HC 96	5·19	70	0·1065	0·519	0·200	0·0274	1·04	5·41
HC 97	0·529	70	2·590	0·605	6·15	0·424	22·03	11·65
HC 99	4·48	70	0·1473	0·542	0·291	0·0375	1·43	6·40
HC 100	3·54	10	0·2040	0·437	0·324	0·0632	2·42	8·56
HC 106	3·55	150	0·2720	0·519	0·523	0·0559	2·15	7·65
HC 123	0·452	150	2·880	0·568	6·31	0·410	22·16	10·02
HC 130	3·53	70	0·2430	0·393	0·378	0·0442	1·70	5·99
HC 137	0·251	70	5·050	0·620	12·61	0·681	45·63	11·45
HC 152	5·79	70	0·0811	0·585	0·172	0·0249	0·95	5·48
HC 155	0·972	10	1·085	0·349	1·59	0·164	6·70	6·51
HC 151	3·50	70	0·2465	0·376	0·374	0·0438	1·67	5·84

* All crystals were grown on a ⟨110⟩ axis with the exception of HC 151 which was grown on a ⟨321⟩ axis.

to K_sG_s and, therefore, G_L becomes smaller. The above pre-supposes that the other parameters in the equation are independent of the growth conditions. However both k and m are functions of concentration.

TRUMBORE *et al.*[11] have determined experimentally the dependence of k on temperature for the binary system Ga–Ge and their results are replotted as k vs. melt concentration (C_L) in Fig. 3 and a smooth curve has been drawn through the points. THURMOND and KOWALCHIK[12] have redetermined the liquidus curve of the Ga–Ge system and have found that its behaviour is near ideal. Employing the assumption that the heat capacities of solid and liquid germanium are equal, then the slope of the liquidus line is given by[1]

$$m = -\tfrac{1}{100}\frac{LT^{\circ 2}R}{\{L-T^{\circ}R\ln y\}^2 y} \qquad (6)$$

where $y = (100-C_L)/100$, T^0 = melting temperature of pure germanium and R = gas constant. Using equation (6) and values of k from Fig. 3 the parameter $p^1 = -m(1-k)/k$ has been calculated and is plotted against C_L in Fig. 4.

In Fig. 5 the experimental data have been plotted in the form $\ln p^1C_s$ vs. $\ln f$, the appropriate values of p^1 being obtained from Fig. 4. The graph has the form predicted by equation (1) (compare Fig. 3 in Ref. 1). In calculating $C_s = k_{\text{eff}}C_L$, the appropriate value of k (that corresponding to the value of C_L) was obtained from Fig. 3 and substituted in the equation (4).

The crystal rotation rate was 70 rev/min for all but four crystals; two crystals (HC 106 and HC 123) were grown at 150 rev/min and two

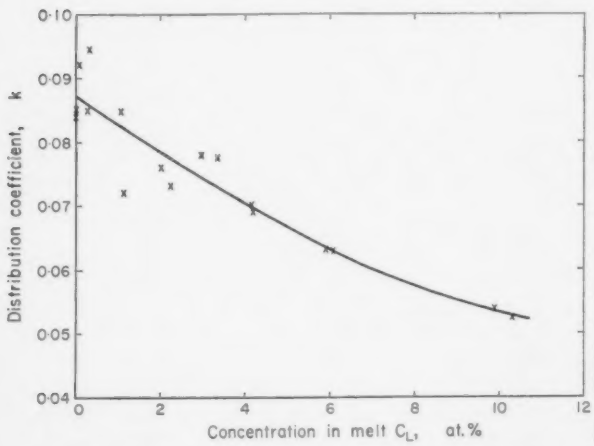


FIG. 3. Dependence of distribution coefficient on melt concentration: gallium in germanium. (Data of TRUMBORE *et al.*[11] replotted.)

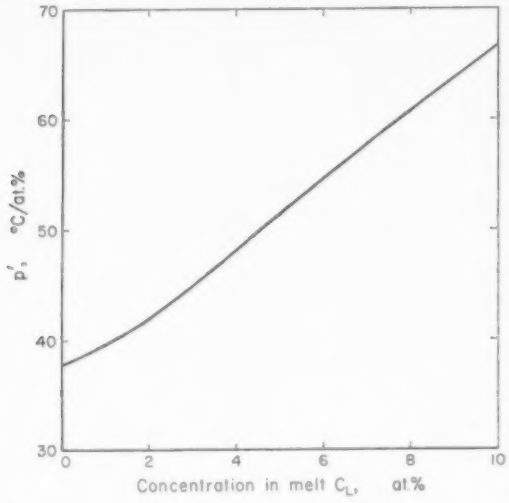


FIG. 4. Dependence of $p^1 = [-m(1-k)]/k$ on melt concentration: gallium in germanium.

(HC 100 and HC 155) at 10 rev/min. The experimental points for all the crystals, irrespective of rotation rate, lie fairly close to the smooth curve in Fig. 5. The radial uniformity of the gallium in crystals rotated at only 10 rev/min is poor: the gallium concentration is highest in the centre of the growing crystal and therefore cell structure forms first in the centre. In crystals grown at the fastest rates (crystals HC 96 and HC 152), cell formation could be distinguished at the periphery at an earlier stage than that at which the crystal was wholly cellular. For the above crystals, the criterion that corrugations indicated the onset of cellular growth was not adequate and therefore the crystals were first sectioned longitudinally and etched to delineate the position at which cell formation occurred in the bulk of the crystal.

The theory of BURTON *et al.*[2] does not account for the radial non-uniformity of solute concentration which is observed in the slowly rotated crystals. The probable reason is that the BURTON *et al.*[2] theory is derived for the case of a planar disk rotating in an infinite melt, whereas, in practice, the melt is contained in a crucible of linear dimensions of the order of twice the radius of the crystal. We have found that the radial uniformity is adequate provided that the rotation rate is greater than 50 rev/min. In our experiments, the crystal radius was 0·9–1·0 cm; the radius of the melt at the commencement of growth was 2·25 cm and $\sim$ 2·0 cm at the point at which cellular growth commenced. The profile of the crucible approximates to a paraboloid of revolution, and the solid–liquid interface was approximately planar.

Fig. 5 shows that, provided the rotation rate is sufficiently large to give a radially uniform concentration of gallium, the rotation rate has no significant effect upon the onset of cellular growth for rotation rates up to 150 rev/min (boundary layer thickness down to 0·007 cm). If the results are expressed in terms of C_L instead of C_s, then the rotation rate does influence the onset of cellular growth because it produces a change in k_{eff}.

Calculated concentrations (Table 1, col. 7), were checked by Hall and resistivity measurements on specimens cut immediately above the cellular region from eight of the crystals, and the values were found to agree, in all cases, to within $\pm$ 15 per cent. Limits of $\pm$ 15 per cent have therefore been drawn on all experimental points in Figs. 5 and 6.

If we now accept the validity of the theory together with the assumption that a fixed critical value of $(dS/dx)_{x=0}$ is required to initiate cell formation, then one would like to calculate this value of $(dS/dx)_x^{(Cr)}{}_0$. The calculation involves a knowledge of G_L. G_L is the temperature gradient in the melt at the interface and, in a stirred melt, the temperature gradient falls with distance (x) from the interface for distances $x > \delta$. Since $0 < x < \delta$ is too small a region in which to measure the temperature gradient directly, G_L must be determined indirectly.

Rewriting equation (1),

$$p^1 C_s f = -D\frac{L\rho}{K_L}f + D\left\{\left(\frac{dS}{dx}\right)_{x=0}^{(Cr)} + \frac{K_s G_s}{K_L}\right\} \tag{7}$$

we see that a graph of $p^1 C_s f$ vs. f should be linear with a slope of $DL\rho/K_L$. The experimental data are plotted in this form in Fig. 6. The two lines drawn are the probable limits for the slope. Using a value of $L = 8100$ cal/g atom (GREINER[13]) and $\rho = 5{\cdot}4$ g/cm^3 we obtain $K_L/D = 3{\cdot}7$–$8{\cdot}1 \times 10^2$ cal cm^{-3} deg^{-1}. BURTON and SLICHTER[14] and BURTON *et al.*[15] give two values for D for gallium in

germanium—namely 10^{-4}[14] and $7{\cdot}5 \times 10^{-5}$[15] cm^2 sec^{-1} respectively. Our Hall measurements on crystals grown at fast rates indicate that the value $D = 10^{-4}\,cm^2\,sec^{-1}$ is the more appropriate. Accordingly we have adopted this value. Hence the range of K_L values is 0·037–0·081 cal cm^{-1} sec^{-1} deg^{-1}. Since the electrical conductivity of molten germanium at the melting point is known, $1{\cdot}56 \times 10^4\,\Omega^{-1}\,cm^{-1}$ (DOMENICALI[16]), a rough estimate for K_L can be obtained using the Lorentz number. This gives a value of $K_L = 0{\cdot}112$ cal cm^{-1} sec^{-1} deg^{-1}, which is somewhat greater than that obtained here. No experimentally determined value for K_L could be found in the literature. The value of K_s for degenerately doped germanium has been determined experimentally by ABELES[17] to be 0·06 cal cm^{-1} sec^{-1} deg^{-1}. (This value is obtained by extrapolating data from ~ 800°C to the melting point.)

From the two intercepts on the ordinate of Fig. 6 and using ABELES'[17] value for K_s and the K_L value obtained from the slope of Fig. 6, it is found that

$$\left(\frac{dS}{dx}\right)_{x=0}^{(Cr)} = 142 - 1{\cdot}62G_s(°C/cm)$$

or

$$102 - 0{\cdot}74G_s(°C/cm) \qquad (8)$$

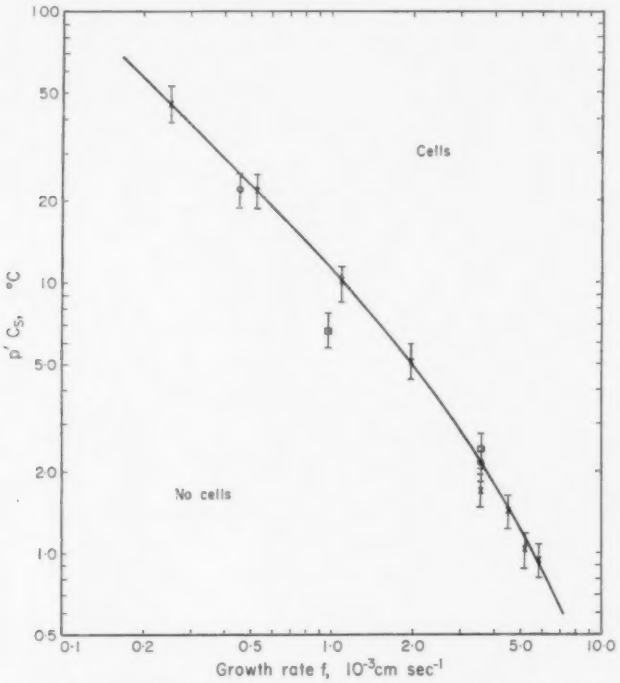


FIG. 5. Graph of $p^1 C_s$ for which cellular growth commenced as a function of growth rate f. (Crystal rotation rates: □ 10 rev/min; × 70 rev/min; ○ 150 rev/min.)

The temperature gradient in the solid (G_s) was determined experimentally by inserting three thermocouples into a crystal and melting back the crystal almost to the lowest of the three thermocouples, then regrowing the crystal at a known rate and simultaneously recording temperature and time. Since the gallium concentration was always adjusted so that cellular growth commenced when between 34·9 per cent and 60·5 per cent of the crystal had grown (see Table 1, col. 5), and since it would be expected that the value of G_s would depend upon the amount of crystal grown, we

determined the value of G_g when approximately 50 per cent of the crystal had been grown. The value of G_g, obtained by extrapolating the temperature–time curve to the interface, was

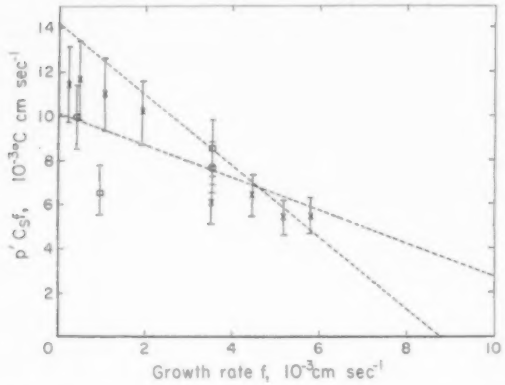


FIG. 6. Graph of $p^1C_s f$ for which cellular growth commenced as a function of growth rate f. (Crystal rotation rates: □ 10 rev/min; ×70 rev/min; ○ 150 rev/min.)

95 ± 15°C/cm. The pull rate used in this experiment was 10^{-3} in/sec with no crystal rotation: the value of G_g is unlikely to depend, to a first order, on either pull rate or rotation rate within the range employed in these experiments. Therefore the values of $(dS/dx)^{(Cr)}_{x=0}$, from equations (8) are −12 and +32°C/cm respectively. The accuracy of these values is poor since the value of $(dS/dx)^{(Cr)}_{x=0}$ is obtained as the difference of two large numbers one of which is uncertain owing to the uncertainty in the values of the thermal conductivities. One can only conclude that $(dS/dx)^{(Cr)}_{x=0}$ must be small (<70°C/cm) or zero.

4. METALLOGRAPHIC EXAMINATION OF THE CRYSTALS

A. *The* {110} *cellular interface*

Several crystals were rapidly withdrawn from the melt soon after corrugations had formed. A magnified portion of a cellular interface is shown in Fig. 7. The cells are very elongated, the long axis being the ⟨110⟩ direction in the plane of the interface. The usual droplet of melt adhered to the centre of the interface as the crystal was withdrawn from the melt, and it can be seen in Fig. 2, that the cellular structure persists into this region. However, only the periphery of the interface was examined where it could be assumed that the macro-structure was characteristic of the interface during steady growth. The highly doped germanium is wetted by its melt more readily than is lightly or undoped germanium. For this reason it has not been possible to examine the cellular interface on the very heavily doped crystals.

The geometrical configurations of the cells were examined using an optical goniometer. Two small {111} facets (arrowed) can be seen at the periphery of the interface in Fig. 2. The {111} facets form one face of the cell at the edge, but the faces of all other cells are less steeply inclined to the (110) surface; the inclination is variable and lies between 5° and 10°. The cell faces frequently exhibit a platelet structure and may be vicinal, i.e. made up of {111} steps. The cell widths are of the order of 0·1–0·2 mm. In the vicinity of the {111} facets, the elongated cells are somewhat broken up. The faces which make up the ends of the elongated cells are inclined at steeper angles to the plane of the interface (~ 24)°, but cannot be identified conclusively with any low-index planes in the germanium lattice.

B. *Delineation of cell boundaries by etching*

The boundaries between the cells have been delineated in the bulk of the crystals by cutting the crystals on {110} and {111} planes and etching. In examining the crystals, three types of surface were used:

(1) The (110) surface parallel to the plane of the interface [(110) para.]. Here the etchant shows the cell boundaries but no dislocations.

(2) The two vertical {111} planes, normal to the plane of the interface ({111} vert.). Here the etchant shows the cell boundaries and dislocation etch pits.

(3) The two {111} planes which are obliquely inclined (at 35° 16′) to the plane of the interface ({111} obl.). Again the etchant shows both cell walls and dislocation etch pits.

The etchant (CUMMEROW and CHERRY[9]), used either hot or cold, was found to show up the cell walls, where they intersect the surface, as dark lines on the polished face. Fig. 8 is a photomicrograph of a (110) para. surface and shows the cell walls with their long axis in the ⟨110⟩ direction.

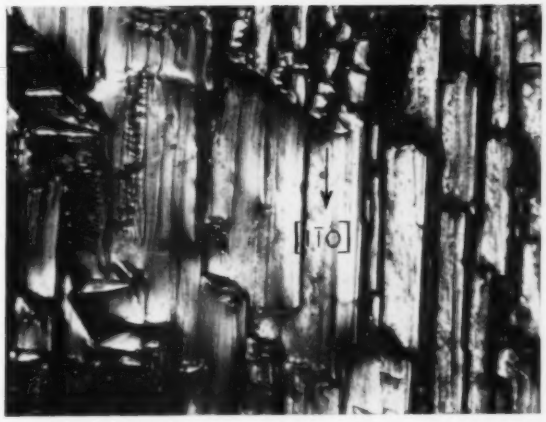


FIG. 7. Portion of a {110} cellular interface. Magnification ×34.

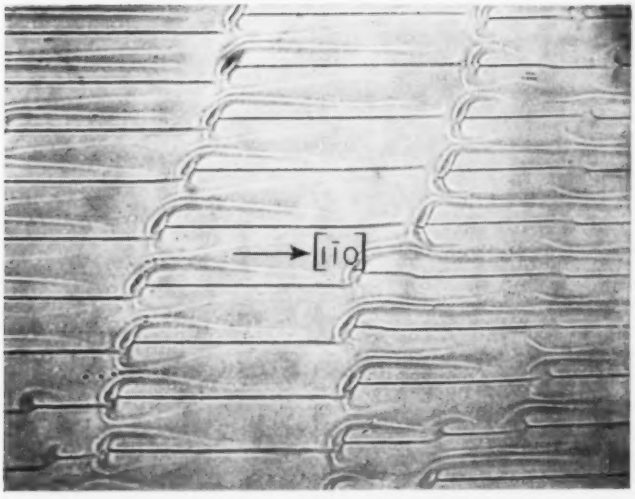


FIG. 8. Regular cell walls on a (110) surface parallel to the plane of the interface delineated by etching. Magnification ×80.

[*facing p.* 148

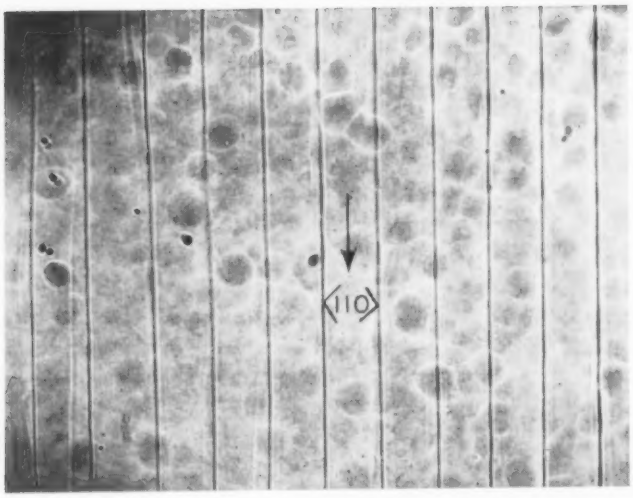


FIG. 9. Regular cell walls on a {111} surface perpendicular to the plane of the interface delineated by etching. Dark circular areas are dislocation pits which are slightly out of focus. Magnification ×62.

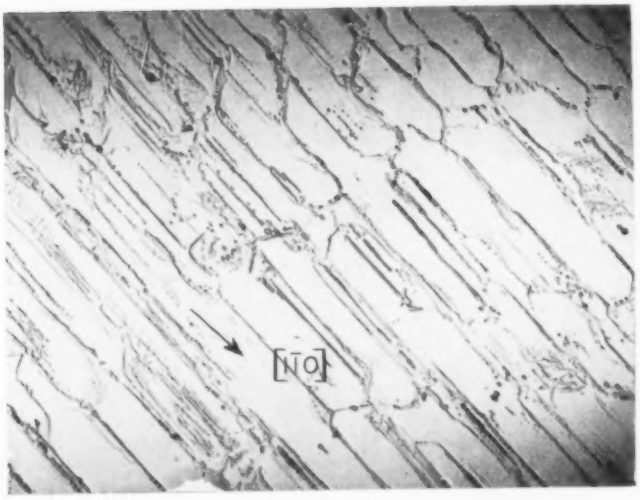


FIG. 10. Irregular cell walls on (110) surface parallel to the plane of the interface. Magnification ×9·6.

The visibility of the cell walls is improved by using the technique of narrow-pencil high-intensity illumination and moving the surface slightly out of focus. The branching at the ends of each cell is also visible. Fig. 9 is a photomicrograph of a {111} vert. face of a crystal taken under similar conditions and shows how the boundaries between the cells are propagated down the crystal. The boundaries can be traced across the common edge of these two faces. Fig. 9 shows that the cells are not in fact macro-steps which sweep laterally across the interface but that the cells grow more or less independently of each other, straight down the crystal. The bounding surfaces between the cells are therefore approximately {100} planes.

The regular cell structures described above are characteristic of the early stages of cellular development. At a more advanced stage, but still with the crystal essentially single, the boundaries between the cells are delineated by irregular closed loops as shown by the photomicrograph of a (110) para. surface (Fig. 10).

Examination of a {111} obl. surface, cut in the irregular cell region, showed clusters of dislocations associated with the enclosed irregular areas. We have found no evidence of any dislocation arrays associated with the early stages of cell development.

The commencement of *regular* cells correlates with the commencement of corrugations and a cellular interface, and this is the criterion adopted in connexion with the quantitative experiments described in this paper.

It is proposed to describe the features of the cellular structure in more detail in a subsequent paper.

C. *Lamellar twinning*

Within the limits of resolution of the optical microscope, no lamellar twinning in the pre-cellular region of carefully grown crystals could be detected. This is in contradiction to the observations of BOLLING *et al.*,[7] that lamellar twinning preceded cell formation. Only in crystals containing lineage have we observed lamellar twinning prior to cell formation.

D. *Distribution of gallium*

It is to be expected that the cell walls contain an excess concentration of gallium and indeed, in crystals grown under extreme conditions of constitutional supercooling, occlusions of free gallium are present.

5. DISCUSSION

A. *The quantitative results*

The conclusion reached, in Section 3, that the critical gradient of constitutional supercooling $(dS/dx)^{(Cr)}_{x=0}$ necessary to initiate cell formation was small or zero is not in agreement with the findings of BOLLING *et al.*[7] These authors state that a concentration of gallium in the germanium twenty times that expected was required to produce cells. In their experiment they obtained a cellular structure when the gallium concentration reached approximately 0·2 at. per cent, whereas, on the basis of the theory of TILLER *et al.*[18] (see below), cellular structure would have been expected with only 0·01 at. per cent gallium. Therefore,

$$\left(\frac{dT_E}{dx}\right)^{(Cr)}_{x=0} = 20G_L \text{ and } \left(\frac{dS}{dx}\right)^{(Cr)}_{x=0} = 19G_L$$

They do not state the value of G_L which they employed but it can be seen that even modest values of G_L (say 20°C/cm) imply large values of $(dS/dx)^{(Cr)}_{x=0}$ (380°C/cm).

The equation of TILLER *et al.*[18] for the onset of constitutional supercooling in an unstirred melt is

$$\frac{G_L}{f} < -\frac{m(1-k)}{Dk}C_L \qquad (9)$$

Equations (1) and (9) are equivalent if

(a) $$\left(\frac{dS}{dx}\right)_{x=0} = \left(\frac{dS}{dx}\right)^{(Cr)}_{x=0} = 0$$

and

(b) $$C_s = C_L \text{ (i.e. } k_{eff} = 1)$$

Equation (9) has been verified experimentally for a number of metal alloys (TILLER and RUTTER,[5] WALTON *et al.*[4] and PLASKETT and WINEGARD[19]). The supposition, that $k_{eff} = 1$, is valid provided that natural convection does not perturb the impurity distribution ahead of the solid–liquid interface which is set up by diffusion. WAGNER[20] has shown that, for a solid–liquid interface lying in a vertical plane, this condition holds provided that f is large and k is near unity.

Possible reasons for the discrepancy in

$(dS/dx)_{x=0}^{(Cr)}$ between the present values and those of BOLLING *et al.*[7] are firstly a difference in the value of D employed (BOLLING *et al.*,[7] do not state the value which they used) and secondly that, in the experiment of BOLLING *et al.*,[7] k_{ett} may not have been equal to unity because of some convective mixing. The use of stirring by crystal rotation in the present experiments reduced the effects of convection and enabled us to calculate accurately the value of k_{ett}.

However, there is a further point: BOLLING *et al.*[7] grew their crystals on a ⟨321⟩ axis whereas, in the experiments described in this paper, the crystals were grown on a ⟨110⟩ axis. To verify that the discrepancy was not due to the difference in orientation, a crystal (HC 151: see Table 1) was grown on a ⟨321⟩ axis in the usual way. It can be seen, comparing HC 151 with its equivalent ⟨110⟩ crystal, HC 130, in Table 1, that the conditions under which cellular growth commences are not different, within the experimental error, from the ⟨110⟩ orientation.

B. *Review*

Several workers have reported gross substructures and impurity inhomogeneities, including the presence of a second phase, in their crystals which are consistent, as is shown below, with their being due to constitutional supercooling (although this has not been consistently appreciated).

MORTIMER[21] and also ALEKSEEVA and ELISEEV[22] noted substructures in pulled crystals of bismuth-doped germanium. MORTIMER[21] found that noticeable flaws appeared in the crystal at about 40 per cent from the top where the melt concentration was ~ 1·3 per cent bismuth, and the ingot became a random polycrystal at about 75 per cent from the top where the bismuth concentration was ~ 3·2 per cent. In Ref. 21 there is a photograph of a transverse section of a crystal, etched in a silver nitrate etchant, on which are deep grooves where excess bismuth had been etched out. The orientation of the crystals is not stated but the grooves are undoubtedly cell walls. ALEKSEEVA and ELISEEV[22] grew ⟨110⟩ and ⟨111⟩ oriented crystals and examined cut longitudinal and transverse {111} faces respectively which had been etched in a $K_3[F(CN)_6]+KOH$ etchant to reveal dislocation etch pits. These authors found that when a concentration of bismuth of 2×10^{16} cm^{-3} in the solid was reached, a sudden increase, by a few orders of magnitude, in dislocation density occurred. Simultaneously with the increase in dislocation density, wavy convex bands appeared on the external surface of the crystals and extended to the bottom of the crystals. It would seem highly probable that these bands were the corrugations of our terminology, and that the increase in dislocation density was associated with the later stages of cell formation. ALEKSEEVA and ELISEEV[22] found that the same concentration of antimony in the solid did not produce these effects. This result is consistent with equation (1) since the distribution coefficient for antimony is greater than that for bismuth ($k_{Sb} = 0{\cdot}003$, $k_{Bi} = 4{\cdot}5\times10^{-5}$); the value of $(dT_E/dx)_{x=0}$ is therefore smaller for antimony. ALEKSEEVA and ELISEEV[22] suggested that the sudden increase in dislocation density was connected with reaching the solubility limit of bismuth in germanium, leading to the incorporation of a second phase. However, they found that, by lowering the growth rate, the concentration of bismuth corresponding to the increase in dislocation density was increased somewhat.

Similar results were obtained by TRUMBORE[23] for germanium crystals pulled from tin-doped melts. TRUMBORE[23] found that, with a pull rate of 3 cm/hr and an unspecified tin concentration, the crystals contained numerous patches of occluded tin about 0·05 mm in diameter. With a pull rate of 0·5 cm/hr, little or no evidence of occlusions was obtained.

TYLER and WOODBURY,[24] TYLER *et al.*[25] and WOODBURY and TYLER[26] report the abrupt onset of a "lineage" structure on the free surface of ⟨100⟩ oriented germanium monocrystals doped with iron, manganese or cobalt. TYLER and WOODBURY[24] demonstrated, using radioactive iron, that the onset of lineage is accompanied by the incorporation of relatively large amounts of iron in pockets in the growing crystals. It was found[25,26] that under the same growth conditions approximately the same concentrations *in the liquid* of manganese or cobalt were required to produce the "lineage" structure, which is consistent with equation (1). The "lineage" obviously corresponds to our corrugations. However these authors report that, by reducing the growth rate from 6 cm/hr to 1·5 cm/hr (a factor of 4), they were able to increase the manganese content in the solid by roughly

ten times. This result is not consistent with equation (1) unless the change in growth rate was accompanied by a large change in temperature gradient in the melt.

ELLIS[27] shows a photograph of an etched section of a ⟨100⟩ oriented melt-grown germanium –(15 per cent) silicon crystal which shows rumpled growth striae. Another figure shows regular cell walls (analogous to our Fig. 8). ELLIS[27] describes this form of growth as equilibrium growth, stating that the interface is composed of pyramidal equilibrium projections. ELLIS[27] also indicates that some of the dislocations lie in the boundaries between the projections, though the photograph is one of regular cell walls with no well-developed dislocation structure.

ROSI[28] describes a hexagonal lineage structure which developed on pure germanium crystals grown at fast rates and also on nickel-doped germanium at ordinary growth rates, and compares the structure with the equilibrium structure described by ELLIS.[27]

TRUMBORE[29] has reported finding small precipitates of germanium arsenide in heavily arsenic-doped germanium crystals. It seems probable that these precipitates form in the cell walls during cooling of the crystal owing to the retrograde solid solubility exhibited by arsenic in germanium. This suggestion is supported by our own observations on cell formation in arsenic-doped germanium.

In Table 2, the relevant experimental growth data, where given, are listed from the literature cited above, and the gradient of the equilibrium liquidus line at the interface corresponding to the point at which the described substructure formed is calculated from the equation[1]

$$\left(\frac{dT_E}{dx}\right)_{x=0} = -\frac{mfC_s(1-k)}{Dk}$$

The appropriate values of m and k are obtained from the reviews of THURMOND and KOWALCHIK[12] and TRUMBORE,[30] respectively. A value of 10^{-4} cm^2 sec^{-1} is used for the diffusion coefficient in all cases. Typically, for liquid metals D lies in the range 10^{-4}–10^{-5} cm^2 sec^{-1} (FRENKEL[31]). If the smaller value of 10^{-5} cm^2 sec^{-1} was appropriate, the values of $(dT_E/dx)_{x=0}$ in Table 2 would be multiplied by a factor of 10.

It will be noted that the values of $(dT_E/dx)_{x=0}$ are all of the same order and are smaller by approximately a factor of 2–5 from those determined experimentally by the present authors. It may be that the diffusion coefficients are slightly smaller than the value of 10^{-4} cm^2 sec^{-1} employed here. A further point is that the pull rate (f^1) rather than the true growth rate (f) has been used in the calculations. (With our crystal pulling equipment f/f^1 is typically 1·3–1·5.)

Since the onset of constitutional supercooling leads to the incorporation of the impurity in higher concentrations than predicted by the normal freeze equation, it is important to know whether or not the data of TRUMBORE *et al.*[11] for the dependence of k on concentration, which has been used in Section 3, was obtained from crystals grown under conditions which avoided the occurrence of constitutional supercooling. TRUMBORE *et al.*[11] produced their crystals both by pulling and by the thermal-gradient technique. The data for the range of concentrations in which we are interested (0–12·6 at. per cent) was obtained from pulled crystals. The crystals were grown in a ⟨111⟩ direction at a pull rate of 0·5 cm/hr with a rotation rate of 144 rev/min. The seed crystals used were of the same diameter as the final crystal. At the slow growth rate and with the high rotation rate, $k_{\text{eff}} \to k$ and hence the slope of the equilibrium liquidus line is given approximately by

$$\left(\frac{dT_E}{dx}\right)_{x=0} = -\frac{mfC_L(1-k)}{D}$$

and is not sensitive to the actual value of k, provided $k \ll 1$. The maximum concentration in the melt which TRUMBORE *et al.*[11] employed was 10·32 at. per cent. Taking $m = 3{\cdot}75$°C/ per cent, $D = 10^{-4}$ cm^2 sec^{-1} and $(1-k) = 1$, then

$$\left(\frac{dT_E}{dx}\right)_{x=0} = 54\text{°C/cm}$$

which is probably slightly less than G_L and therefore constitutional supercooling was probably not present. The possibility of anisotropic segregation cannot be ignored (HULME and MULLIN[32]), and the dependence of k on C_L may be different for growth in the ⟨110⟩ direction and in the ⟨111⟩ direction. HALL[33] found no orientation

Table 2

Ref.	Doping element	Distribution coefficient k	m °C/at.%	Pull rate f[1]	Rotation rate ω (rev/min)	C_s (at.%)	C_L	$(dT_E/dx)_{x=0}$ (°C/cm)	Orientation	Comments
21	Bi	$4{\cdot}5\times10^{-5}$	3·7	$\frac{3}{4}$ in/hr	Not stated	5×10^{-5}	~1·3%	21·8	Not stated	Etched transverse section shows cell walls
22	Bi	$4{\cdot}5\times10^{-5}$	3·7	1 mm/min	Not stated	$4{\cdot}4\times10^{-5}$		60·9	⟨110⟩ ⟨111⟩	Noted large increase in dislocation density
23	Sn	2×10^{-2}	3·56	0·5 cm/hr	60		~5 at.%	~24	⟨100⟩	Patches of occluded Sn ~0·05 mm in diameter
24	Fe	$\sim3\times10^{-5}$	3·7*	~2 in/hr	~100		~0·3 at. %	~16	⟨100⟩	Fe occlusions; Co shows same effect at approx. same C_L†
26	Mn	1×10^{-6}	3·7*	1·5 cm/hr	Not stated	$\sim2{\cdot}2\times10^{-6}$		~34	⟨100⟩	"Lineage" markings on crystal surface

* No experimental data available. Value quoted obtained from $m = T_0^2R/L$ (ideal dilute solution).
† Ref. 25.

dependence for gallium in germanium but DIKHOFF[34] reports that gallium exhibits a small "facet effect" in germanium, the ratio k (on the facet) to k (off the facet) being 0·85. Thus the data of TRUMBORE *et al.*[11] may be slightly, but should not be seriously, in error.

6. CONCLUSIONS

Making the assumption that the regular cellular structure forms at a fixed critical gradient of supercooling $(dS/dx)^{(Cr)}_{x=0}$, the theory predicting the onset of constitutional supercooling in a stirred melt has been verified for a wide range of concentrations and growth rates for the case of the growth of ⟨110⟩ oriented germanium crystals from gallium-doped melts. Further, it is concluded that the value of $(dS/dx)^{(Cr)}_{x=0}$ is probably zero and anyway not greater than 70°C/cm. We have been unable to specify the limits of $(dS/dx)^{(Cr)}_{x=0}$ more precisely because of the uncertainty in the value of K_S/K_L used in the calculation of G_L. The conclusions that $(dS/dx)^{(Cr)}_{x=0} \sim 0$ was substantiated, within the experimental limits, for a crystal grown on a ⟨321⟩ axis. The latter result is not in agreement with the findings of BOLLING *et al.*[7] who studied the growth of the same alloy system from an unstirred melt. The reason for the discrepancy cannot be firmly deduced because BOLLING *et al.*[7] do not give sufficient quantitative data.

The advantage gained by the use of a stirred melt in a study of the effects of constitutional supercooling is that stirring imposes a solute distribution in the melt which can be accurately calculated and which is not seriously affected by convective mixing even at the slowest growth rates. By contrast, with unstirred melts, the solute distribution in the melt at slow growth rates may be dominated by convection effects and the solute distribution cannot be readily calculated.

Evidence of cellular structure has been obtained in three ways: (a) by the corrugations on the free surface of the crystal, (b) by rapidly withdrawing the crystal from the melt during growth to reveal the cellular growth surface and (c) by etching on (110) and {111} surfaces of the crystal to reveal the boundaries between the cells. Two types of cell wall, designated regular and irregular, have been identified by etching, and it has been noted that a large increase in dislocation density is associated with the irregular cell walls. As might be expected, the cells show a much more pronounced crystallographic character than do the cells on the growth surface of a metal single crystal.[6]

The results of some other workers who have reported various substructures in germanium crystals grown from doped melts have been analysed and it is concluded that the results indicate, both qualitatively and quantitatively, that the substructures were produced as a result of the existence of constitutional supercooling in the melt during growth. The results relate to a variety of solute elements in germanium (namely, bismuth, tin, iron, manganese and cobalt) and the analysis suggests that the efficacy of a dopant in producing constitutional supercooling is specified by the magnitude of $p = [-m(1-k)]/Dk$. The theory shows that the production of germanium uniformly doped with solutes which have very small distribution coefficients is difficult, because these solutes readily produce a zone of constitutional supercooling ahead of the growing crystal.

One may conclude that the concept of constitutional supercooling is of general applicability to the growth of germanium, and probably other semiconductors, from heavily doped melts. The theory derived in Ref. 1 enables the upper limit to the growth rate, whilst avoiding constitutional supercooling, to be predicted.

The concept of constitutional supercooling is also applicable to the growth of compound semiconductors from non-stoichiometric melts. A cellular structure can be produced on the interface of an indium antimonide crystal by growing it either from an indium-rich or an antimony-rich melt (MULLIN[35]). The concept should be of value when considering the maximum rate of growth which can be achieved when growing semiconductor compounds where the equilibrium vapour pressure of one of the constituents at the melting point is too high for it to be convenient to grow crystals from stoichiometric melts (e.g. gallium phosphide, indium phosphide).

The deliberate use of constitutional supercooling during growth may prove a useful technique for fundamental studies of the growth of semiconductors from the melt and from solution and should supplement the information which

has been obtained on the characteristics of growth under conditions of thermal supercooling (dendritic and polyhedral growth[36–38]).

Acknowledgements—The authors are indebted to Mr. G. W. GREEN for making the Hall coefficient measurements, to Dr. J. B. MULLIN for valuable discussions and to the Controller, H. M. Stationery Office, for permission to publish.

REFERENCES

1. D. T. J. HURLE, *Solid-State Electron.* **3,** 37 (1961).
2. J. A. BURTON, R. C. PRIM and W. P. SLICHTER, *J. Chem. Phys.* **21,** 1987 (1953).
3. J. W. RUTTER and B. CHALMERS, *Canad. J. Phys.* **31,** 15 (1953).
4. D. WALTON, W. A. TILLER, J. W. RUTTER and W. C. WINEGARD, *J. Metals, N.Y.* **7,** 1023 (1955).
5. W. A. TILLER and J. W. RUTTER, *Canad. J. Phys.* **34,** 96 (1956).
6. C. ELBAUM, *Progress in Metal Physics* Vol. 8, p. 203. Pergamon Press, London (1959).
7. G. F. BOLLING, W. A. TILLER, and J. W. RUTTER, *Canad. J. Phys.* **34,** 234 (1956).
8. L. ESAKI, *Phys. Rev.* **109,** 603 (1958).
9. R. L. CUMMEROW and A. R. CHERRY, *Phys. Rev. Letters* **3,** 367 (1959).
10. W. G. PFANN, *Zone Melting* Chap. 2. Wiley, New York (1958).
11. F. A. TRUMBORE, E. M. PORBANSKY and A. A. TARTAGLIA *J. Phys. Chem. Solids* **11,** 239 (1959).
12. C. D. THURMOND and M. KOWALCHIK, *Bell Syst. Tech. J.* **39,** 169 (1960).
13. E. S. GREINER, *J. Metals, N.Y.* **4,** 1044 (1952).
14. J. A. BURTON and W. P. SLICHTER, *Transistor Technology* Vol. 1, Chaps. 5 and 6. Van Nostrand, New York (1958).
15. J. A. BURTON, E. D. KOLB, W. P. SLICHTER and J. D. STRUTHERS, *J. Chem. Phys.* **21,** 1991 (1953).
16. C. A. DOMENICALI, *J. Appl. Phys.* **28,** 749 (1957).
17. B. ABELES, *J. Phys. Chem. Solids* **8,** 340 (1959).
18. W. A. TILLER, K. A. JACKSON, J. W. RUTTER and B. CHALMERS, *Acta Met.* **1,** 428 (1953).
19. T. S. PLASKETT and W. C. WINEGARD, *Canad. J. Phys.* **37,** 1555 (1959).
20. C. WAGNER, *J. Metals. N.Y.* **6,** 154 (1954).
21. G. MORTIMER, *J. Electrochem. Soc.* **105,** 739 (1958).
22. V. G. ALEKSEEVA and P. G. ELISEEV, *Soviet Phys. Solid State* **1,** 1195 (1960).
23. F. A. TRUMBORE, *J. Electrochem. Soc.* **103,** 597 (1956).
24. W. W. TYLER and H. H. WOODBURY, *Phys. Rev.* **96,** 874 (1954).
25. W. W. TYLER, R. NEWMAN and H. H. WOODBURY, *Phys. Rev.* **97,** 669 (1955).
26. H. H. WOODBURY and W. W. TYLER, *Phys. Rev.* **100,** 659 (1955).
27. S. G. ELLIS, *J. Appl. Phys.* **26,** 1140 (1955).
28. F. D. ROSI, *RCA Rev.* **19,** 349 (1958).
29. F. A. TRUMBORE, *Semiconductor Symposium of the Electrochemical Society, May,* 1959 (unpublished).
30. F. A. TRUMBORE, *Bell Syst. Tech. J.* **39,** 205 (1960).
31. J. FRENKEL, *Kinetic Theory of Liquids* p. 201. Oxford University Press (1946).
32. K. F. HULME and J. B. MULLIN, *Phil. Mag.* **4,** 1286 (1959).
33. R. N. HALL, *Phys. Rev.* **88,** 139 (1952).
34. J. A. M. DIKHOFF, *Solid-State Electron.* **1,** 202 (1960).
35. J. B. MULLIN. Private communication (1961).
36. R. S. WAGNER, *Acta Met.* **8,** 57 (1960).
37. A. I. BENNETT and R. L. LONGINI, *Phys. Rev.* **116,** 53 (1959).
38. D. R. HAMILTON and R. G. SEIDENSTICKER, *J. Appl. Phys.* **31,** 1165 (1960).

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 155–158. Printed in Great Britain

NOTE

Secondary multiplication in silicon*

(*Received* 13 *March* 1961)

IN A recent paper by SHOCKLEY[1] a simple model was used to calculate the secondary multiplication coefficient α in silicon and germanium. The theory was shown to be in good agreement with experimental results for electric fields ranging from 2×10^5 to 5×10^5 V/cm. Only two scattering processes for carriers are considered in SHOCKLEY's paper: (*a*) generation of Raman phonons of energy ϵ_r and mean-free-path l_r and (*b*) ionization or electron–hole pair production with threshold carrier energy ϵ_i and mean-free-path l_i. Of these four characteristic constants, ϵ_r is determined independently from neutron-scattering data while the other three constants are adjusted to fit a number of different experimental data. It is shown in SHOCKLEY's paper that the three adjustable constants permit fitting six pieces of experimental data in four independent experiments: (1) quantum yield for photons with $1 < h\nu < 5$ eV, (2) ionization coefficient α as a function of the electric field E, (3) energy per pair for ionization by high energy particles and (4) spectra of hole–electron recombination in microplasmas.

It is the purpose of this note to show that a refinement of SHOCKLEY's calculation of α leads, for some of the adjustable constants, to values which are quite different from those quoted by SHOCKLEY. This is especially so for the ratio, $r \equiv l_i/l_r$, between the two mean-free-paths.

Most of the experimental results available[2–4] cover the so-called low-field region in SHOCKLEY's notation, i.e. such fields which give rise to a mean energy $\langle \epsilon \rangle$ of the hot electrons which is small compared to ϵ_i. For this case SHOCKLEY derives the expression

$$\alpha(E) = \frac{qE}{r\epsilon_r}\exp(-\epsilon_i/qEl_r) \qquad (1)$$

when $r \gg 1$.†

The derivation of this expression is based on the following model. Consider a carrier scattered to an energy less than ϵ_r. It will reach the energy ϵ_i if it travels a distance ϵ_i/qE without scattering. The probability of its doing this is $\exp(-\epsilon_i/qEl_r)$. Having reached ϵ_i the probability of ionizing in the first scattering is $(1/r) = l_r/l_i$. The total probability of ionizing per try in this single-stage process is then

$$P_1 = (1/r)\exp(-\epsilon_i/qEl_r) \qquad (2)$$

In the stationary state, the energy qE gained from the field when drifting 1 cm must equal the sum of the energies lost to phonons and, in ionizing collisions,

$$qE = \alpha\epsilon_i + \alpha_r\epsilon_r \qquad (3)$$

where α_r is the number of phonon collisions taking place when the electron travels 1 cm in the field direction.

The total number of collisions per centimeter (= total number of "tries" per centimeter) times the probability of ionizing gives the ionizing coefficient

$$\alpha = (\alpha_r + \alpha)P_1 \qquad (4)$$

Inserting (2) and (3) into (4) yields

$$\alpha = \frac{qE}{r\epsilon_r}\exp(-\epsilon_i/qEl_r) \times\left[1+\left(\frac{\epsilon_i}{\epsilon_r}-1\right)\frac{1}{r}\exp(-\epsilon_i/qEl_r)\right]^{-1} \qquad (5)$$

* This work was sponsored by Signal Corps Contract DA36(039) SC–85339.

† The expressions in Ref. (1) as well as in this note require this approximation. Only trivial changes have to be made when this approximation is not used.

In the low-field case, (5) approximates into the expression (1).

The above treatment considers only the ionizations arising from electrons scattered to an energy less than ϵ_r, then gaining the necessary energy from the field without any scatterings and subsequently ionizing in the first attempt. Ionization, due to carriers which reach, for example, 0·5 ϵ_i and then are scattered by a phonon and subsequently gain the necessary energy, is neglected. These two-stage processes are claimed to be of negligible importance because a carrier which has been scattered into a random direction with energy 0·5 ϵ_i must on the average travel farther to reach ϵ_i than a carrier starting at rest. Another process, which is not considered, is the one where an electron reaches an energy larger than ϵ_i without scattering, then has a Raman scattering and subsequently ionizes. It will be shown in the following that taking into account these multi-stage processes gives rise to important corrections to the expression (1).

Let us consider an electron starting at $x = 0$ with the velocity $v = 0$. The probability that this electron goes a distance x_1 without scattering and then has a phonon scattering within dx_1 (at $\epsilon < \epsilon_i$) is

$$\exp(-x_1/l_r)\, dx_1/l_r \tag{6}$$

The phonon scattering mechanism gives rise to an isotropic velocity distribution so that the probability of the electron being scattered an angle θ from the direction of the field is $(\sin\theta\, d\theta)/2$. If the electron energy *after* the phonon scattering is ϵ_1, the electron must gain $\epsilon_i - \epsilon_1$ from the field before it can ionize. Having gained this energy, the probability of ionizing in the first scattering is $1/r$. The total probability of this two-stage process is thus

$$P_2(\epsilon_1 < \epsilon_i) = \int_0^{\epsilon_i+\epsilon_r} \int_0^{\pi} \times \exp(-x_1/l_r)\frac{dx_1}{l_r} \times \exp[-s(\theta, x_1)/l_r]\frac{1}{r}\frac{\sin\theta\, d\theta}{2} \tag{7}$$

where $s(\theta, x_1)$ is the distance an electron which is scattered at $x = x_1$ into an angle θ with the field must travel in order to gain the energy $\epsilon_i - \epsilon_1$:

$$s(\theta, x_1) = \frac{\epsilon_1}{qE}\left\{\sqrt{\frac{\epsilon_i}{\epsilon_1}}\sqrt{\left(\frac{\epsilon_i}{\epsilon_1} - \sin^2\theta\right)} - \cos\theta + \sin^2\theta \ln\left[\frac{\sqrt{\epsilon_i/\epsilon_1} + \sqrt{\left(\frac{\epsilon_i}{\epsilon_1} - \sin^2\theta\right)}}{1+\cos\theta}\right]\right\} \tag{8}$$

Fig. 1 is a plot of the total distance of travel $x_1 + s(\theta, x_1)$ normalized by the distance $s_0 = \epsilon_i/qE$ (corresponding to no collisions) as a function of ϵ_1/ϵ_i. It may be noted that the deviation from s_0 is less than 20 per cent for all values of ϵ_1/ϵ_i when $\theta < 45°$.

The two-stage process with $\epsilon_1 \geqslant \epsilon_i$ can be treated analogously. In this case some of the electrons which scatter in angles between $\pi/2$ and π may be decelerated by the field to energies less than ϵ_i and thus have the probability zero instead of $1/r$ of ionizing in their second scattering. A conservative estimate of the correction to (1) is obtained by discounting all electrons decelerated to energies less than ϵ_i, thereby neglecting the contribution from those which are later brought up again by the field to $\epsilon \geqslant \epsilon_i$. We thus obtain

$$P_2(\epsilon_1 \geqslant \epsilon_i) = \int_{(\epsilon_i+\epsilon_r)/qE}^{\infty} \exp(-x_1/l_r)\left(1 - \frac{1}{r}\right)\frac{1}{r} \times \left\{ \int_0^{\theta_{\min}} \tfrac{1}{2}\sin\theta\, d\theta + \int_{\theta_{\min}}^{\pi} \tfrac{1}{2}(1 - \exp[-s(\theta, x_1)/l_r]\sin\theta\, d\theta)\right\}\frac{dx_1}{l_r} \tag{9}$$

where $s(\theta, x_1)$ is again the distance an electron must travel in order to gain the energy $\epsilon_i - \epsilon_1$ ($\epsilon_i - \epsilon_1$ being a negative number in this case), while $\theta_{\min}$ is the minimum scattering angle ($> \pi/2$) from which the electron can reach energies less than ϵ_i. The fraction of electrons scattered at $x = x_1$ which can reach energies less than ϵ_i is rapidly decreasing when x_1 increases from $(\epsilon_i + \epsilon_r)/qE$. Using expressions (7–9) one finds that the number of electrons with $\epsilon_1 > \epsilon_i$ which are decelerated

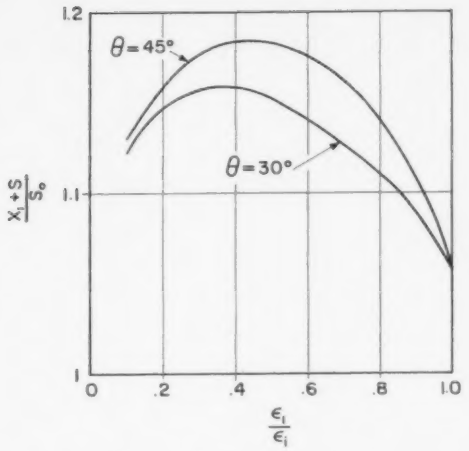


FIG. 1. Total travel distances $x_1+s(\theta, x_1)$ corresponding to an energy gain of ϵ_i with one phonon scattering normalized by the distance $s_0 = \epsilon_i/qE$ as a function of ϵ_1/ϵ_i.

below ϵ_i is more than compensated for by the number of electrons with $\epsilon_i-\epsilon_r \leqslant \epsilon_1 \leqslant \epsilon_i$ which are accelerated to energies larger than ϵ_i. A conservative estimate of the total probability $[P_2(\epsilon_1 < \epsilon_i)+P_2(\epsilon_1 \geqslant \epsilon_i)]$ of the two-stage process can therefore be found by lowering the upper limit of the first integral in (7) to the value ϵ_i, while at the same time changing $P_2(\epsilon_1 \geqslant \epsilon_i)$ to

$$P_2'(\epsilon_1 \geqslant \epsilon_i) = \int_{(\epsilon_i+\epsilon_r)/qE}^{\infty} \frac{r-1}{r^2} \exp(-x_1/l_r)\, d\left(\frac{x_1}{l_r}\right)$$

$$= \frac{r-1}{r} \exp(-\epsilon_r/qEl_r)\left[\frac{1}{r}\exp(-\epsilon_i/qEl_r)\right] \quad (10)$$

A three-stage process starting with two phonon scatterings can be treated analogously. In order that the energy ϵ_2 of the electron immediately after the second phonon scattering is still larger than ϵ_i we must have $x_1 \geqslant (\epsilon_i+2\epsilon_r)/qE$. The expression for $P_3'(\epsilon_2 \geqslant \epsilon_i)$ is then

$$P_3'(\epsilon_2 \geqslant \epsilon_i) = \int_{(\epsilon_i+2\epsilon_r)/qE}^{\infty} \left(1-\frac{1}{r}\right)^2 \frac{1}{r}$$

$$\times \exp(-x_1/l_r)\, d\left(\frac{x_1}{l_r}\right) = \left(\frac{r-1}{r}\right)^2$$

$$\times \exp(-2\epsilon_r/qEl_r)\left[\frac{1}{r}\exp(-\epsilon_i/qEl_r)\right] \quad (11)$$

In order to find the contributions from the two-stage processes with $\epsilon_1 < \epsilon_i-\epsilon_r$, one has to solve (7) with the corrected upper limit of the first integral and with $s(\theta, x_1)$ given by (8). This expression is not directly integrable but a rather crude calculation suffices to establish the following lower limit

$$P_2(\epsilon_1 < \epsilon_i-\epsilon_r) \geqslant 0{\cdot}5$$

$$\times \exp(-\epsilon_r/qEl_r)\left[\frac{1}{r}\exp(-\epsilon_i/qEl_r)\right] \quad (12)$$

It follows from the above that a conservative correction of SHOCKLEY's theory can be found by

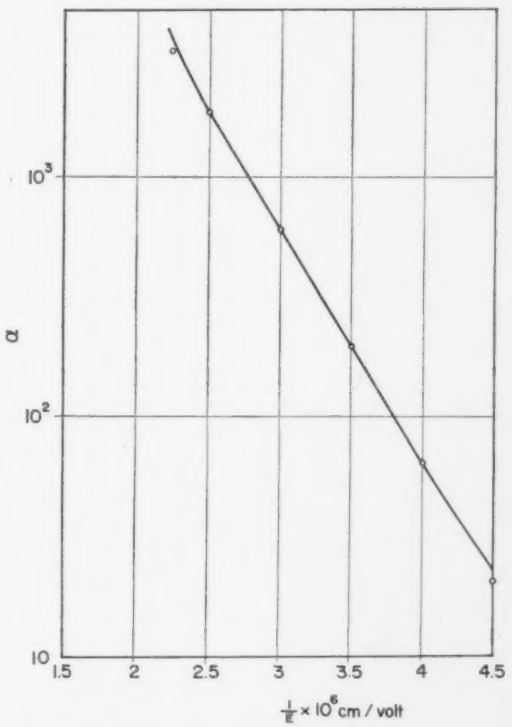


FIG. 2. Theoretical curve as calculated from expression (14) for the ionization coefficient α as a function of electric field. ○ = experimental points.[2]

replacing P_1 given in (2) by P_{tot}; where

$$P_{tot} = P_1 + P_2(\epsilon_1 < \epsilon_l - \epsilon_r) + \sum_{n=2}^{\infty} P'_n(\epsilon_{n-1} \geqslant \epsilon_l) \quad (13)$$

When (13) is inserted instead of P_1 in the expression for α, one gets

$$\alpha = \frac{qE}{r\epsilon_r} \times \exp(-\epsilon_i/qEl_r)\left\{\frac{1}{1-(r-l/r)\exp(-\epsilon_r/qEl_r)} + 0{\cdot}5\exp(-\epsilon_r/qEl_r)\right\}. \quad (14)$$

Using the values for r, l_r and ϵ_r for silicon given in SHOCKLEY's paper[1]

$$r = 18; \qquad l_r = 50\,\text{Å}; \qquad \epsilon_r = 0{\cdot}063\ \text{eV} \quad (15)$$

the value of the correction factor in (14) is

2·3 at $E = 2\times10^5$ V/cm and

4·1 at $E = 5\times10^5$ V/cm

The expression (14) for α can be fitted to the experimental data[2] for electrons in silicon using the same values for ϵ_l and ϵ_r if a value of $r \simeq 200$ is used instead of $r = 18$ and a value of $l_r = 64$ instead of $l_r = 50$ (see Fig. 2). With this higher value of r it is, however, no longer possible to fit SHOCKLEY's theory to the experimental results for the photon quantum yield and the energy per pair for ionization by high-energy particles. It therefore seems necessary to reconsider the theoretical and experimental basis of some of the data treated in SHOCKLEY's paper.

Stanford Electronics Laboratories J. J. MOLL
Stanford University N. I. MEYER*
Stanford California

REFERENCES

1. W. SHOCKLEY, *Solid-State Electron.* **2,** 35 (1961).
2. A. G. CHYNOWETH, *J. Appl. Phys.* **31,** 1161 (1960).
3. S. L. MILLER, *Phys. Rev.* **105,** 1246 (1957).
4. A. G. CHYNOWETH, *Phys. Rev.* **109,** 1537 (1958).

* On leave from the Technical University of Denmark, Copenhagen.

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 159–166. Printed in Great Britain

INDIUM ANTIMONIDE TRANSISTORS*†

H. L. HENNEKE

Central Research Laboratories, Texas Instruments Inc., Dallas 22, Texas

(*Received* 1 *March* 1961; *in revised form* 1 *June* 1961)

Abstract—The design, fabrication, and electrical characteristics of an *n–p–n* indium antimonide transistor which operates at 77°K are discussed. An analysis of the expected high-frequency performance is presented and a comparison made to a *p–n–p* germanium transistor. Because of its high electron mobility, the indium antimonide transistor should have a base transit time 1/26 that of a similar germanium transistor. Although these ultimate capabilities have not been approached, transistors with f_t as high as 300 Mc/s and β's of 500 have been observed. Switching speeds approaching those of the best state-of-the-art germanium transistors have been measured.

Résumé—Le dimensionnement, la fabrication, et les caractéristiques d'un transistor *n–p–n* à l'antimoniure d'indium opérant à 77°K sont discutes dans cet article. Une analyse du rendement à haute fréquence attendu est présenté et une comparaison faite avec un transistor au germanium. Grâce à sa forte mobilité d'électrons, le transistor à l'antimoniure d'indium devrait avoir un parcours électronique 26 fois plus rapide que celui d'un transistor similaire au germanium. Malgré que ces capacités extrêmes n'ont pas été atteintes, des transistors ayant des f_t aussi élevées que 300 Mc/s et des βs de 500 ont été observés. Des vitesses de commutation approchant les meilleurs transistors au germanium ont été mesurées.

Zusammenfassung—Struktur, Herstellung und die elektrischen Charakteristiken eines *n–p–n*-Indiumantimonid-Transistors, der bei 77°K arbeitet, werden diskutiert. Der zu erwartende Hochfrequenzbereich wird analysiert und mit dem eines *p–n–p*-Germanium-Transistors verglichen. Wegen seiner beträchtlichen Elektronenbeweglichkeit sollte die Laufzeit in der Basis bei diesem Transistor nur 1/26 von der eines ähnlichen Germanium-Transistors sein. Obwohl man die äussersten Leistungen noch nicht erreicht hat, liessen sich Frequenzen f_t bis zu 300 MHz und β-Werte bis 500 beobachten. Die gemessenen Schaltgeschwindigkeiten näherten sich denen der besten existierenden Germanium–Transistoren.

1. INTRODUCTION

BECAUSE indium antimonide has the highest electron mobility of all III–IV semiconducting compounds, it is a most promising starting material for either a microwave or a high-speed switching transistor.

The primary difficulty in attempting to construct bipolar transistors from compound semiconductors like indium antimonide arises from the low carrier-lifetime values. This short lifetime results in short carrier diffusion lengths and requires relatively narrow base regions to obtain sufficient current gain. In fact, the base regions required are comparable in width to those achieved on the most recent high-frequency germanium and silicon transistors.

To investigate the high-frequency performance expected for an indium antimonide transistor, a theoretical analysis was made to relate the physical parameters of such a transistor to its high-frequency response. An *n–p–n* structure was chosen for the model.

2. NOTATION

C_c Collector transition capacity.
C_e Emitter transition capacity.
D_n Diffusion constant for electrons.

* The research reported in this paper has been supported in part by the Bureau of Naval Weapons and has been performed under the technical direction of the Naval Ordnance Laboratory, White Oak.

† A limited account of this work was presented at the Solid State Device Research Conference in Pittsburgh, Pennsylvania, June, 1960.

D_p Diffusion constant for holes.
f_α α cut-off frequency.
$f_{\max\ osc}$ Maximum frequency of oscillation.
f_t Frequency at which $|h_{fe}|$ is unity.
h_{fe} High-frequency common-emitter current gain.
J_e Emitter current density.
k Boltzmann's constant.
n A constant between 1 and 6, depending on the base-layer impurity distribution.
N_c Donor impurity concentration in the collector.
N_e Donor impurity concentration in the emitter.
P_b Acceptor concentration in the base region.
P_c Acceptor concentration in the collector region.
P_e Acceptor concentration in the emitter region.
q Electron charge.
r_b' Base spreading resistance.
t_b Base transit time.
t_c Collector depletion layer transit time.
t_e Emitter charging time.
t_{ec} Emitter-to-collector transit time for carriers.
T Temperature in °K.
$v_{sc\ lim}$ Lattice-scattering-limited maximum drift velocity for carriers in the collector barrier region.
W Base width.
X Collector depletion-layer width.
β_0 Large signal β.
μ_n Electron mobility.
μ_p Hole mobility.
τ_s Storage time constant.

3. ANALYSIS OF THE CAPABILITIES OF THE INDIUM ANTIMONIDE TRANSISTOR

The frequency performance of a junction transistor may be characterized fully by the parameters f_α, the frequency at which the common-base current gain has fallen off 3 dB, r_b', the base series resistance, and C_c, the collector junction capacity. These are related by the expression[1]

$$(\text{power gain})^{1/2}(\text{bandwidth}) = f_{\max\ osc} = \frac{1}{4\pi}\left(\frac{1}{r_b' C_c t_{ec}}\right)^{1/2} \quad (1)$$

where t_{ec} is the emitter-to-collector transit time for carriers and is equal to the sum of the emitter charging time t_e, the base transit time t_b, and the collector depletion layer drift time t_c.

The complete expression for $t_{ec}(f_\alpha = 1/2\pi t_{ec})$ is[2]

$$t_{ec} = \frac{kT}{qJ_e}C_e + \frac{W^2}{nD_n} + \frac{X}{2v_{sc\ lim}} \quad (2)$$

To obtain a value for t_{ec}, we must reduce each term of this equation to measurable physical quantities. If this is done for a drift-transistor structure with graded base-impurity distribution, a comparison may be made between the theoretical limits and capabilities of an indium antimonide and a germanium transistor.

Assume a germanium p–n–p transistor with $P_c = 4\times10^{15}$ cm^{-3}, $P_e = 10^{19}$ cm^{-3} and an average $N_b = 10^{17}$ cm^{-3}. Compare this to an indium antimonide n–p–n transistor with $N_c = 4\times10^{14}$ cm^{-3}, $N_e = 10^{18}$ cm^{-3} and an average $P_b = 10^{16}$ cm^{-3}. Further, assume the same physical structure, base width and impurity variation for the two transistors. A reasonable value for the drift mobility of minority carriers in the base of the germanium transistor is 1000 cm^2/V-sec for $T = 300$°K, so that $D_p = \mu_p kT/q = 26$ cm^2/sec. For the indium antimonide transistor, a value of 100,000 cm^2/V-sec is reasonable for μ_n at $T = 78$°K; so that $D_n = \mu_n kT/q = 670$ cm^2/sec. The lattice-scattering-limited maximum drift velocity is approximately 8×10^6 cm/sec for germanium[3] and may be greater than 2×10^7 cm/sec for indium antimonide.[4]

From equation (2) the following comparisons may be made for the two transistors:

t_e—InSb less than Ge by a factor of 6
t_b—InSb less than Ge by a factor of 26
t_c—InSb less than Ge by a factor of 5

Since the average resistivity of the base layer is the same for the two transistors, r_b' will be approximately the same. If the lower operating voltage of the indium antimonide transistor is considered, C_c is also about the same for the two cases.

At this point it is important to note that if the emitter changing time t_e and the collector transit time t_c are negligible so that base transit time t_b dominates, the α-cut-off frequency will actually

be higher than the value $1/2\pi t_b$ for either the germanium or the indium antimonide transistor. KROEMER[5] has shown that other factors must be considered in relating base transit time to a cut-off frequency resulting from this transit time. His result, a more accurate expression for f_α of a drift transistor, is

$$f_\alpha \simeq \frac{W}{\pi D^{1/2}} \frac{1}{(2t_b)^{3/2}} \tag{3}$$

Substituting for the value of t_b,

$$f_\alpha \simeq \frac{D}{2\pi W^2} \left(\frac{\ln P_0/P_b(W)}{2^{1/3}} \right)^{3/2} \tag{4}$$

The conclusion from these calculations is that the indium antimonide transistor may be expected to have an α-cut-off frequency of from 5 to greater than 26 times that of a geometrically similar germanium transistor. The actual value will depend on which transit time dominates.

4. METHOD OF FABRICATION

To utilize properly the high electron mobility of indium antimonide, our effort has been concentrated on developing an *n–p–n* transistor. Because of the proved usefulness of solid-state vapor diffusion for forming narrow base regions, this technique was used to form the *p*-type base layer. The emitter region could best be formed by alloying because: (1) A transistor capable of high-frequency operation must have a small emitter area to reduce junction capacity. Techniques employed in silicon technology, which allow impurity-diffusion selective masking to reduce emitter area, are not yet developed for indium antimonide. (2) For the highest emitter injection efficiency, an abrupt emitter-base junction (such as those obtained by alloying) is preferable to the graded junction produced by diffusion. (3) Preliminary diffusion studies of normal group VI donors indicate difficulty in obtaining the high concentrations necessary for good emitter efficiency.

The starting material is single crystal *n*-type indium antimonide grown by the Teal–Little technique at the Central Research Laboratories of Texas Instruments. The impurity concentration is about $4 \times 10^{14}\,\text{cm}^{-3}$. The electron mobility is greater than 400,000 cm^2/V-sec, as determined by Hall measurements at 77°K. This material is cut in slices 0·040 in. thick on the ⟨111⟩ plane and is lapped to a thickness of about 0·020 in. with a fine abrasive. The slices are then optically polished. Thorough washings in triple-deionized water follow each step. The *p*-type base region is formed by vacuum diffusion in a sealed quartz ampule.

Although a number of impurity-diffusion sources were investigated in elemental and dilute alloy form, none was completely satisfactory. Zinc,[6] which was investigated most thoroughly, was found to be unsatisfactory for several reasons. (1) The surface concentration from an elemental source is too high, approaching the limit of zinc solubility in indium antimonide, $2 \times 10^{20}\,\text{cm}^{-2}$. (2) The diffusion coefficient is extremely concentration-dependent. Lowering the surface concentration by reducing zinc vapor pressure ("diluting" the zinc in indium) reduces the diffusion coefficient to an impractically small value. (3) The diffusing zinc does not follow a complementary error-function distribution. The diffusion profile contains a break where concentration falls off rapidly with distance. Such a profile makes accurate control of the base width under the alloyed emitter contact difficult.

Other acceptor impurities—cadmium, magnesium, manganese and mercury—were also investigated. Magnesium shows promise, but no completely satisfactory diffusant has yet been found.

After a *p*-type diffused layer about 10μ thick is obtained, one side of the slice is lapped down to remove this layer and to expose the original *n*-type crystal. The slices are then diced into smaller wafers about 0·040 by 0·040 in., and mesas are masked and etched on each wafer. The mesas are about 0·010 in. in diameter. The collector connection is formed by soldering the lapped side of each wafer with pure tin to a TO-18 header. The ohmic base connection is made by alloying a 0·002 in. pellet of 98 per cent indium/2 per cent cadmium alloy onto the diffused layer in a strip-heater micro-furnace under helium atmosphere. The emitter contact is formed by similarly alloying a 0·002 in. pellet of an indium–gallium–tellurium alloy onto the *p*-type layer. This contact is not allowed to penetrate the base layer and forms the emitter junction of the *n–p–n* structure. Gold wires 0·0007 in. in diameter connect the header

to these contacts. The transistors are then given a short clean-up etch to reduce surface leakage currents. Fig. 1 is a sketch of the completed diffused base-alloyed emitter structure.

5. EXPERIMENTAL RESULTS

Fig. 2 is a photograph of the curve-tracer presentation of an indium antimonide transistor operating common base at 77°K. The α of the unit is about 0·975 at a collector current of 10 mA. Fig. 3 shows the characteristics of this transistor, illustrating the ability to operate at very low levels. The α is still about 0·75 at 50 μA collector current. Figs. 4 and 5 illustrate the base–collector and base–emitter diode characteristics of this transistor. These diodes have less reverse current and more nearly diffusion-limited forward current than any indium antimonide diodes previously reported in the literature.

Figs. 6 and 7 are photographs of this same transistor operating common emitter at both high and low current levels. Beta is about 40 at an emitter current of 7 mA.

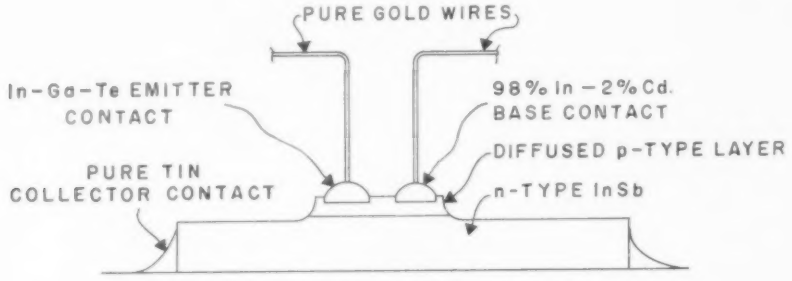


FIG. 1. Structural view of indium antimonide n–p–n mesa transistor.

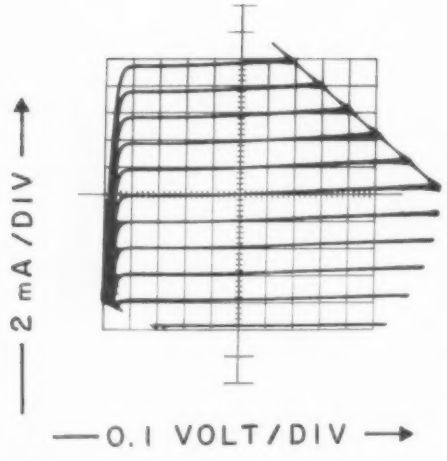


FIG. 2. Indium antimonide transistor characteristics at high current level, common base. Emitter current is 2 mA/step, $\alpha = 0{\cdot}975$.

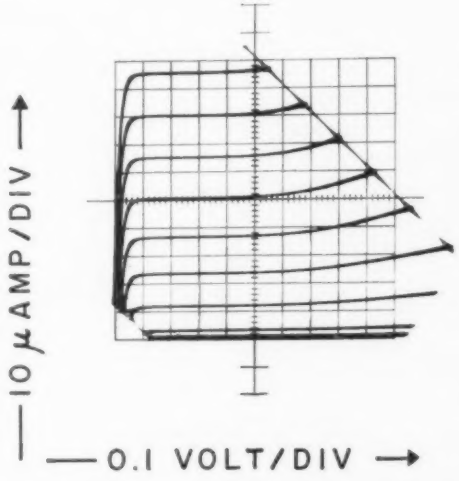


FIG. 3. Indium antimonide transistor characteristics at low current level, common base. Emitter current is 20 μA/step, $\alpha = 0{\cdot}75$.

Although most of the transistors had β's between 10 and 40, a few exhibited much greater current gains. Fig. 8 is a photograph of the characteristics of one of these higher gain transistors. The α of this unit is about 0·994 at 5 mA collector current.

The low breakdown voltage is apparent in all these photographs. At present, the breakdown is thought to be due to avalanche breakdown.

The common-emitter high-frequency current gain h_{fe} was measured by conventional means. Fig. 9 shows the variation of h_{fe} with frequency

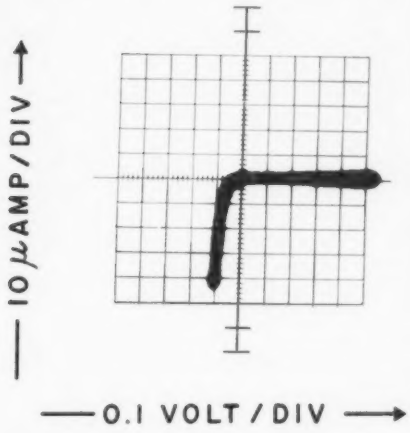


FIG. 4. Indium antimonide transistor, base-collector diode characteristics, emitter open circuited. $I_{CBO}=1\mu A$; BV_{CBO} = 3V at 10 μA.

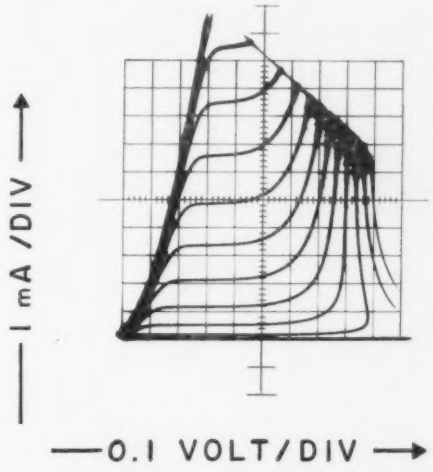


FIG. 6. Indium antimonide transistor characteristics at high current level, common emitter. Base current is 50 μA/step, β = 40.

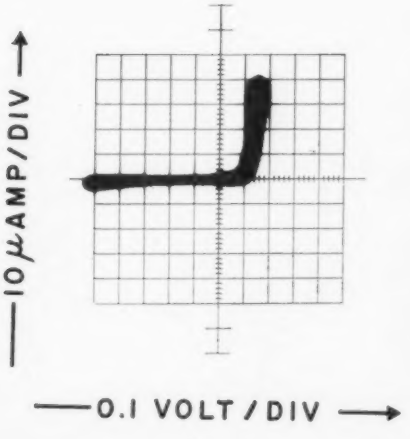


FIG. 5. Indium antimonide transistor, base-emitter diode characteristics, collector open circuited. I_{EBO} = 2 μA; BV_{EBO} = 1 V at 10 μA.

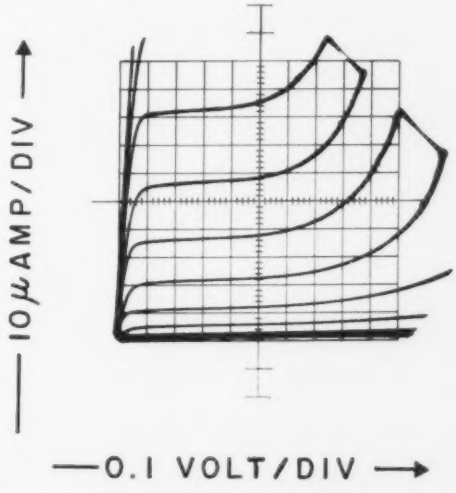


FIG. 7. Indium antimonide transistor characteristics at low current level, common emitter. Base current is 10 μA/step, β = 2.

for one of the better indium antimonide transistors. At the frequency where h_{fe} is unity, f_t is about 160 Mc/s for this transistor.

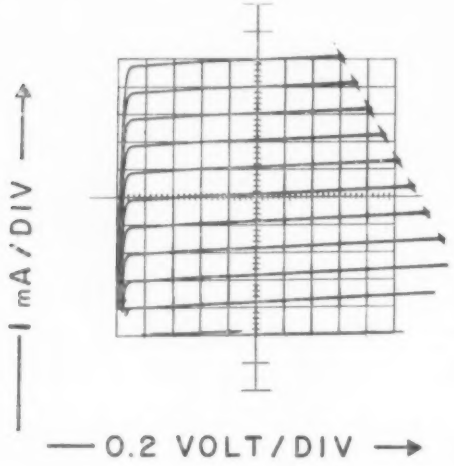


FIG. 8. Indium antimonide transistor characteristics, common base. Emitter current is 1 mA/step, $\alpha = 0{\cdot}994$.

Typical parameters and the best parameters observed during this period of research are tabulated in Table 1.

Table 1. *Indium antimonide transistor parameters at* 77°K

Parameter	Conditions	Typical	Best observed
β	$I_c = 5$ mA	20	500
I_{CBO}	$V_c = 0{\cdot}5$ V	20 μA	<1 μA
I_{EBO}	$V_E = 0{\cdot}3$ V	20 μA	<1 μA
BV_{CBO}	$I_c = 1$ mA	3 V	20 V
BV_{EBO}	$I_E = 1$ mA	0·5 V	2 V
R_{CS}	$I_c = 10$ mA	15 Ω	4 Ω
h_{ib}	$I_E = 5$ mA	20 Ω	7 Ω
h_{ob}	$I_c = 5$ mA	500 μmho	10 μmho
f_t	$I_E = 3$ mA	80 Mc/s	300 Mc/s

Since electron and hole lifetimes in indium antimonide are very low, indium antimonide transistors are advantageous in saturating circuitry. Accordingly, these transistors were compared with commercial transistors designed for fast-switching circuits.

There are two ways to study storage effects: (1) compare storage times for the same amount of overdrive on the base (common-emitter connection), or (2) compare a time constant associated with collector-current decay in the storage mode. The first is widely used. The second may require

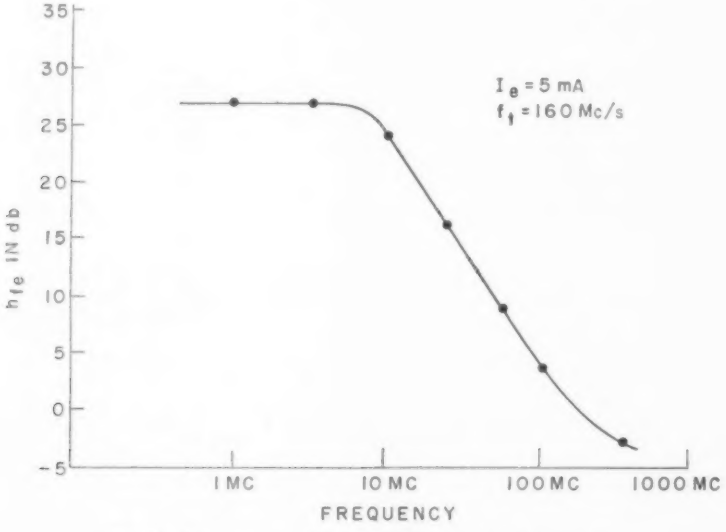


FIG. 9. Common emitter current gain vs. frequency.

some explanation (see Fig. 10). It is assumed[7] that the collector current rises to a level $I_{CM} = \beta_0 I_b$, where β_0 is large-signal β, and I_b is base current. Actually, the collector current is limited to $I_{CS} \simeq V_{cc}/R_L$, where R_L is the load resistor; therefore, storage time is time required for the collector current to fall from I_{CM} to $0{\cdot}9\, I_{CS}$. The storage time constant τ_s governs this fall.

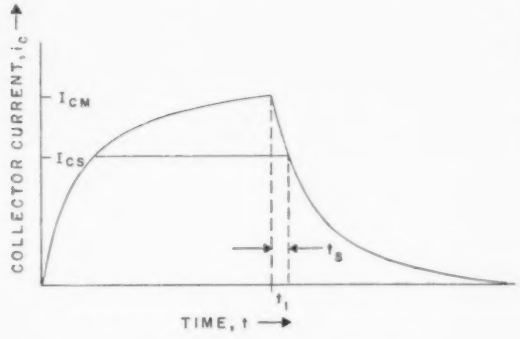


FIG. 10. Collector current wave form.

In Fig. 10, the base current is switched off at t_1, and the storage time is t_s.
For $t \geqslant t_1$,

$$i_c = I_{CM} \exp(-t/\tau_s) \tag{5}$$

Applying end-point values from storage time definitions,

$$0{\cdot}9 I_{CS} = I_{CM} \exp(-t_s/\tau_s) \tag{6}$$

or

$$\frac{0{\cdot}9 V_{cc}}{R_L} = \beta_0 I_b \exp(-t_s/\tau_s) \tag{7}$$

The storage time constant is then

$$\tau_s = \frac{t_s}{\ln[\beta_0 I_b R_L / 0{\cdot}9 V_{cc}]} \tag{8}$$

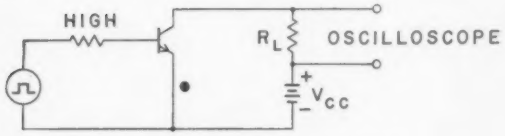


FIG. 11. Storage-time test circuit.

Using the circuit shown in Fig. 11, it is possible to determine I_b, R_L, V_{cc} and t_s. From the Tektronix 575 curve tracer, β_0 is found; thus, τ_s is calculated from (8). This figure of merit has the advantage of relative independence from the degree of saturation and will be used for comparing various transistors. Storage time results for a number of transistors are tabulated in Table 2.

Table 2. Storage-time measurements

Transistor	β_0	I_{cs} (mA)	t_s (mμsec)	I_b (mA)	τ_s (mμsec)
Indium antimonide (77°K)—V_{cc} = 1 V; R_L = 390 Ω; V_b = 10 V					
A	25·00	0·50	190	0·05	40·5
B	2·00	1·40	70	5·00	33·8
C	1·65	1·65	50	5·00	29·2
D	1·78	1·60	90	5·00	49·5
E	3·00	1·50	180	5·00	75·0
Germanium (300°K)—V_{cc} = 1 V; R_L = 390 Ω; V_b = 10 V					
2N501–1	35·00	1·75	160	5·00	34·0
2N501–2	32·00	1·60	260	5·00	55·4
2N501–3	22·80	1·60	170	5·00	39·0
2N240–1	7·00	1·75	230	5·00	74·4
2N240–2	8·75	1·75	270	5·00	81·3

6. CONCLUSIONS

The work reported in this paper has demonstrated the feasibility of fabricating transistors from the semiconductor compound indium antimonide. These transistors exhibit low input impedances and operate at very low power levels. In conventional switching circuits, these transistors have switched in times comparable to the best state-of-the-art germanium transistors. The reasons for the relatively low frequency cut-off compared to the theoretical capabilities are not clearly understood. As the limitations on performance of this new transistor are understood, the research will be reported.

Acknowledgements—The author is grateful to his colleagues in the Device Research Department of Texas

Instruments for many helpful discussions and suggestions. I should also like to acknowledge the early work on indium antimonide devices of R. L. PETRITZ (presently at Texas Instruments Incorporated, Dallas, Texas) of the Naval Ordnance Laboratory. For developing the circuitry and carrying out the switching measurements, I am indebted to R. W. LADE, presently at Marquette University. Special thanks are due Mrs. B. MIMS and Mrs. F. EUBANKS for assistance in fabricating the transistors.

REFERENCES

1. R. L. PRITCHARD, *Frequency Response of Grounded-Base and Grounded-Emitter Transistors* AIEE Winter Meeting, New York, January, 1954.
2. J. M. EARLY, *Proc. I.R.E.* **46,** 1924 (1958).
3. E. J. RYDER, *Phys. Rev.* **90,** 766 (1953).
4. M. GLICKSMAN and M. C. STEELE, *Phys. Rev.* **110,** 1204 (1958).
5. H. KROEMER, *Arch. elektr. Ubertr.* **8,** 363 (1954).
6. D. L. KENDALL and M. E. JONES. To be published.
7. D. E. DEUITCH, *Transistors I.* RCA Laboratories (1956).

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 167–207. Printed in Great Britain

THEORY OF THE ESAKI DIODE FREQUENCY CONVERTER

ROBERT A. PUCEL

Research Division, Raytheon Company, Waltham 54, Massachusetts

(*Received* 14 *April* 1961)

Abstract—The small-signal conversion properties of an Esaki (tunnel) diode are represented by a simple two-port conductance matrix whose elements are certain coefficients of the periodic time-dependent diode conductance produced by the local oscillator (pump). Because of the negative slope in the diode I-V characteristic, arbitrarily high conversion gain is possible when certain conditions are satisfied by these coefficients.

In terms of these coefficients and other diode parameters, expressions are derived for such useful converter properties as the load conditions necessary for circuit stability, the minimum noise figure, the maximum gain-bandwidth product, the necessary conditions for a positive mixer conductance at the radio frequency (r.f.) and intermediate frequency (i.f.) ports, and the pump loading necessary for self-excitation. It is shown that under proper conditions partial noise cancellation can occur because of correlation effects arising from the nonstationarity of the shot-noise process in the pumped diode.

Most of the theoretical results are illustrated by the data obtained from a detailed numerical Fourier analysis applied to an actual high-frequency Esaki diode characteristic. These calculations show that the lowest noise figure and the highest gain-bandwidth product are obtained when the diode is biased in its negative conductance region.

Résumé—Les propriétés de la conversion de faibles signaux d'une diode Esaki (tunnel) sont représentées par une matrice de conductance à 2 termes dont les éléments sont certains coefficients d'une conductance de diode en fonction de temps produite par un oscillateur (pompe). Dû à la pente negative de la caractéristique I–V de la diode, une amplification arbitraire de conversion fort élevée est possible quand certaines conditions sont satisfaites par ces coefficients.

En fonction de ces coefficients et d'autres paramètres de la diode des expressions sont derivées pour d'utiles propriétés de conversion telles que les conditions de charge nécessaires à la stabilité du circuit, le facteur minimum de bruit, le produit maximum gain-largeur de bande, les conditions nécessaires pour une conductance positive de mélangeur aux jointes r.f. et i.f. et la charge 'pompe' nécessaire à l'auto-excitation. Il est démontré que sous certaines conditions appropriées, on peut arriver à une élimination partielle du bruit grâce aux effets de correlation produits par la mobilité du procédé de bruit dans la diode pompée.

La plupart des résultats théoriques sont illustrés par les données obtenues à l'aide d'une analyse numérique détaillée de Fourier appliquée à une caracteristique actuelle de diode Esaki de haute fréquence. Ces calculs démontrent que le facteur de bruit minimum et le produit gain-largeur de bande maximum sont obtenus quand la diode est polarisée dans sa region de conductance négative.

Zusammenfassung—Die Verstärkereigenschaften einer Esaki-(Tunnel)Diode für schwache Signale werden durch eine einfache Leitwert-Matrize mit zwei Zuleitungen dargestellt, deren Elemente gewisse Koeffizienten des von der Periode abhängigen Leitwerts der Diode sind, der durch den lokalen Oszillator (Pumpe) erzeugt wird. Wegen der fallenden I–V-Kennlinie der Diode, ist eine beliebig hohe Verstärkung möglich, wenn diese Koeffizienten gewisse Bedingungen erfüllen.

Mittels dieser Koeffizienten und anderer Diodenparameter werden Ausdrücke für nützliche Verstärkereigenschaften abgeleitet: Belastungsbedingungen für Schaltkreisstabilität, minimale Rauschzahl das maximale Produkt aus Verstärkung und Bandbreite, die erforderlichen Bedingungen für einen positiven Mischungsleitwert an den r.f.- und i.f.-Zuleitungen und die zur Selbsterregung nötige Pumpenbelastung. Es ergibt sich, dass bei geeigneten Bedingungen eine teilweise Beseitigung

des Rauschens eintritt, wegen der Korrelationseffekte, die durch den nicht-stationären Schroteffekt in der aufgepumpten Diode entstehen.

Die theoretischen Ergebnisse werden illustriert, indem eine detaillierte numerische Fourieranalyse an der Kennlinie einer Hochfrequenz–Esaki–Diode ausgeführt wird. Die Berechnungen zeigen, dass die neidrigste Rauschzahl und das höchste Produkt aus Verstärkung und Bandbreite dann erzielt wird, wenn die Diode in dem Gebiet des negativen Leitwerts vorgespannt wird.

INTRODUCTION

NONLINEAR elements which are used for mixing or frequency conversion fall into two broad categories: (a) vacuum tubes and (b) semiconductor diodes. If we exclude variable capacitance semiconductor diodes (varactors), these two categories have one feature in common: namely, that the incremental device conductance (diode) or transconductance (triode and multigride tube) presented to an incoming small signal varies periodically in time in accordance with a larger, locally applied pump signal. As we shall see later, this is the basic mechanism underlying the frequency-conversion process.

The vacuum tube mixer, being frequency limited by transit-time effects, is relegated to applications below 500 to 1 000 Mc/s.[1] On the other hand, the conventional (positive conductance) crystal mixer, because of its negligible transit-time effects, has found widespread use at the higher microwave frequencies. However, being a passive device, it must necessarily exhibit a conversion power loss between the incoming signal and its outgoing, frequency-translated counterpart. Depending on the frequency of the incoming microwave signal, the lowest conversion losses obtained with conventional crystal mixers terminated in a low-noise i.f. stage are in the neighborhood of 4·5 dB and higher. The corresponding noise figures fall in the range from 5·5 to 8 dB.[2] The lack of gain is objectionable, particularly since it is not accompanied by a very low noise figure.

As is well known, many microwave systems employ frequency converters, particularly down-converters (input frequency greater than output frequency). It has not been possible in the past to incorporate high-gain, low-noise down-converters into such systems. Although the introduction of varactor diodes has made possible the realization of such converters under certain conditions,[3] recently reported results indicate that these objectives can also be realized with Esaki diodes, and possibly with less difficulty.

The Esaki diode, unlike the varactor diode, is of the nonlinear conductance type, but is distinguished from the conventional crystal diode by the presence of a region of negative slope in its highly nonlinear current–voltage (I–V) characteristic.[4] Because of this negative conductance region, the Esaki mixer is an active device and hence can exhibit power gain. This possibility was first demonstrated by CHANG *et al.*[5], who reported a down-converter with power gain in excess of 20 dB in the VHF range. Down-converters with gain operating in the microwave range were reported later by others.[6,7] Equally as important is the fact that noise figures as low as 4 dB were obtained in some cases.[5,7]

Unlike other diode mixers, the Esaki mixer can be self-excited since the negative conductance region allows one to operate the diode as its own pump. The only power supply necessary is the bias source.

To date, the theory of Esaki mixers has not been considered in sufficient detail in the literature to indicate the conditions for optimum mixer performance. It was with this need in mind that the theoretical investigation summarized in this paper was made. Applying the theoretical results of our investigation to a numerical example based on an actual high-frequency Esaki diode characteristic, we demonstrate the degree of optimum performance that one can reasonably expect of present-day Esaki diodes, and probably of those in the near future.

NOTATION

I	diode tunnel current
I_n	diode equivalent noise current
I_e, I_z	oppositely directed components of tunnel current
V	diode junction voltage
$J(t)$	sum of tunnel current and displacement current of diode
C	junction capacitance of diode

L	series inductance of diode package
R_b	series resistance of diode
$V_p(t)$	pump component of diode junction voltage
$v(t)$	small-signal component of junction voltage
$I_p(t)$	pump component of total diode current
$i(t)$	signal component of total diode current
$g(V)$	incremental diode conductance
g_k	kth harmonic amplitude of $g[V(t)]$
ω_p	fundamental pump frequency (Mc/s)
ω_1	signal input frequency (Mc/s)
ω_2	signal output frequency (Mc/s)
ω_i	image frequency (Mc/s)
V_1, V_2, V_3 (I_1, I_2, I_3)	complex sinusoidal amplitude of signal input, output and image voltages (currents), respectively
y_{jk}	jkth admittance element of conversion matrix
g_n	equivalent noise conductance
η	ratio of load to source temperature, T_l/T_s
V_0	bias component of pump voltage
$2V_{p1}, (2I_{p1})$	peak amplitude of fundamental component of pump voltage (current)
G	large-signal diode conductance presented to pump (mmhos)
P	pump power (μW)
I_s	complex amplitude of signal input current source
$\Omega_1, \Omega_2, \Omega_3$	resonance frequency of input, output and image ports, respectively
K_0	midband transducer gain of mixer
B	bandwidth of mixer (Mc/s)
c	output–input capacitance ratio
μ	modulation coefficient
$(t-t_k)$	elementary tunnel-current pulse
$\rho(t)$	tunnel emission rate (no./sec)
n_1, n_2, n_3	equivalent shot-noise generators at input, output and image ports
γ	correlation coefficient of n_1 and n_2
γ'	correlation coefficient of n_1 and n_3
I_{nk}	kth harmonic amplitude of $I_n(t)$
F	total noise figure
F_s	shot-noise figure
T	absolute temperature (°K)
T_s	standard source temperature (300°K)

1. DERIVATION OF THE SMALL-SIGNAL CONVERSION MATRIX; CONSIDERATION OF MIXER ACTIVITY AND STABILITY

A typical I–V characteristic of an Esaki diode is sketched in Fig. 1(a). For mixer applications, the segment of the "forward" bias range labeled O–A on the voltage axis, containing the region of negative slope, is of prime interest.* For this range, the diode plus its holder can be represented by the circuit model shown in Fig. 1(b). The variable

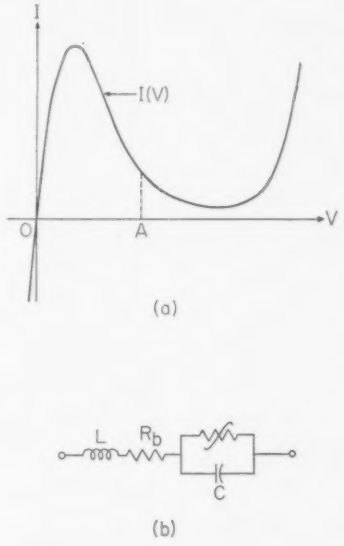


FIG. 1. (a) I–V characteristic, and (b) equivalent circuit of an Esaki diode.

conductance represents the nonlinear I–V characteristic. Since this conductance arises from the quantum-mechanical tunnel effect, it can be considered frequency independent.[4] The capacitor represents the usual junction barrier capacitance. For Esaki diodes this capacitance exhibits only a mild voltage dependence in the region of interest.[8] The series resistor denotes the bulk resistance of the semiconductor material external to the p–n junction and the resistance of the contacts made to the semiconductor.

* The reason for excluding the remainder of the negative-slope region and the valley region will be made later.

The series inductor represents the inductance of the diode leads and of the holder.

For the present we shall asume that the series impedance of L and R_b can be neglected in the frequency range of interest since its inclusion complicates the analysis unnecessarily. We do not wish to imply that this assumption is always justifiable in practice. In a later Section we shall indicate how our analysis can be generalized to include this impedance. Little error is made in assuming C to be voltage independent in the region of interest, though the analysis can be readily extended to include any voltage dependence if desired.

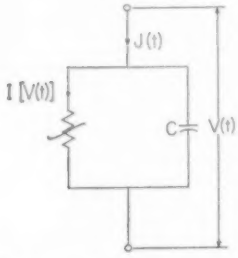


FIG. 2. Simplified equivalent circuit of an Esaki diode.

1.1 *Small-signal conversion matrix*

The time-dependent continuity equation for the simplified diode model (Fig. 2) is

$$J(t) = I[V(t)] + C\frac{\partial V(t)}{\partial t} \tag{1.1}$$

The total instantaneous diode current $J(t)$ consists of a sum of two components: an in-phase conductive part $I[V(t)]$ attributable to the diode tunnel current [Fig. 1(a)] and a quadrature part equal to the displacement current of the barrier capacitance. We suppose that the total impressed voltage is composed of a large pump or local oscillator component $V_p(t)$ and a much smaller signal component $v(t)$, so that we may write

$$V(t) = V_p(t) + v(t) \tag{1.2}$$

Because of this "small-signal" assumption, we may expand the nonlinear tunnel component in a Taylor series about the instantaneous pump operating point $I[V_p(t)]$ and retain only the first two terms. Hence,

$$I[V(t)] = I[V_p(t)] + g(t)v(t) + \ldots \tag{1.3a}$$

where

$$g(t) = g[V_p(t)] \equiv \left.\frac{\partial I(V)}{\partial V}\right|_{V = V_p(t)} \tag{1.3b}$$

may be interpreted as the instantaneous incremental conductance. With this approximation, it follows that the total diode current also may be decomposed into a large pump component $I_p(t)$ and a small signal component $i(t)$, that is,

$$J(t) = I_p(t) + i(t) \tag{1.4}$$

Thus equation (1.1) separates into two equations, one applicable to the large-signal pump components,

$$I_p(t) = I[V_p(t)] + C\frac{\partial V_p(t)}{\partial t} \tag{1.5}$$

and the other to the small signal components,

$$i(t) = g(t)v(t) + C\frac{\partial v(t)}{\partial t} \tag{1.6}$$

Unlike the pump equation, the latter is linear with voltage. This implies that the resulting converter will also be linear. The linearity stems from our small-signal assumption since only terms to first order in $v(t)$ were retained in the Taylor expansion. Observe that although the signal current is coupled to the pump voltage via the time-varying conductance, the pump current is not coupled to the signal voltage. This too is a consequence of the small-signal approximation.

It may be evident by now that frequency conversion occurs via the time-varying conductance. This conductance, which obviously is a periodic function having the period τ of the pump, can be expanded in a Fourier series as follows:

$$g(t) = \sum_{k=-\infty}^{\infty} g_k \exp(jk\omega_p t) \tag{1.7a}$$

$$= g_0 + g_1 \exp(j\omega_p t) + g_1^* \exp(-j\omega_p t)$$
$$+ g_2 \exp(2j\omega_p t) + g_2^* \exp(-2j\omega_p t) + \ldots$$
$$= g_0 + 2|g_1| \cos(\omega_p t + \phi_1)$$
$$+ 2|g_2| \cos(2\omega_p t + \phi_2) + \ldots$$

where

$$g_k = \frac{1}{\tau}\int_0^\tau g(t)\exp(-jk\omega_p t)dt \quad (g_{-k} = g_k^*) \quad (1.7b)$$

Here ω_p represents the fundamental pump frequency ($\omega_p\tau = 2\pi$). The asterisk denotes the complex conjugate.

In general, every g_k is complex except g_0, the time-averaged conductance. However, since the time origin is arbitrary, at least one additional g_k can always be assumed real (actually real and positive) for any general pump voltage. For convenience, from here on we shall choose this to be g_1. Furthermore, if it is possible to express the pump voltage as an even function of time by a suitable choice of the time origin, then *every* g_k can be assumed real. The most general periodic pump voltage satisfying this requirement is of the form

$$V_p(t) = \sum_{n=0}^{\infty} V_{pn}\cos[n(\omega_p t+\phi)] \quad (1.8)$$

where ϕ is a constant. Obviously, a sinusoidal pump voltage consisting of a constant (bias) term plus a fundamental component fulfills this condition. If equation (1.8) holds, then $g(t)$ can be written as

$$g(t) = g_0+2\sum_{k=1}^{\infty} g_k\cos k\omega_p t \quad (1.9)$$

For definiteness let ω_1 denote the frequency of the incoming signal voltage applied to the diode. Because of the periodicity of $g(t)$, components of current are generated at the frequencies $|k\omega_p \pm \omega_1|$, where $k = 0, 1, 2, \ldots$. These currents, in passing through the external circuitry and the junction capacitor in turn develop voltages across the diode at these same frequencies. Suppose that we wish to maintain only two signal voltages across the diode, one at the input frequency ω_1, and one at the desired output frequency ω_2, which is related to ω_1 by some specified multiple of the fundamental pump frequency. To realize this condition we postulate that all undesired voltage components, or "sidebands" as they are often called, are "shorted out" by ideal filters shunting the diode. These filters are assumed to have negligible effect on the desired signal voltages and on the pump voltage. In a later Section we shall discuss this assumption in detail and the consequences when it is not fulfilled.

One possible circuit with which the above conditions can be achieved is shown in Fig. 3.

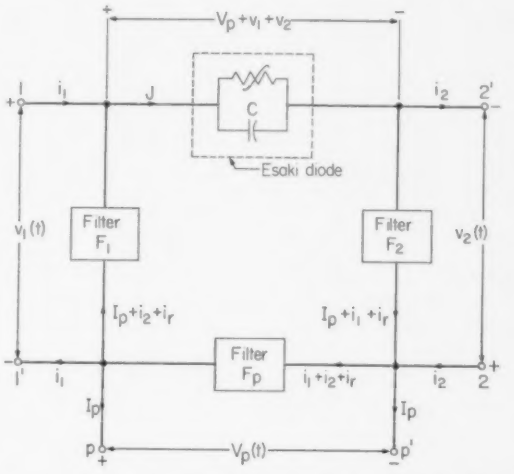


FIG. 3. Model pertinent to derivation of the small-signal conversion matrix.

The signal input and output ports are represented by the terminal pairs 1–1′ and 2–2′, respectively, and the pump port by p–p'. The ideal filter F_1 across the signal input port is assumed to have an infinite or very "large" impedance at ω_1, and zero or a very "small" impedance at all other frequencies, including that of the pump. The filter F_2 across the output port satisfies similar requirements at ω_2. Similarly, the pump filter F_p presents a negligible impedance at the signal and sideband frequencies and a large impedance at the pump freqeuncy ω_p. In practice, these filter requirements can be approximated by simple high-Q parallel resonant circuits. Even tuned circuits may be dispensed with in some cases. For example, if one of the signal frequencies is very low, the corresponding filter can be simulated by a simple capacitor. Furthermore, it is not absolutely essential that the impedance of the pump filter be high at ω_p if sufficient pump voltage is available. However, if this filter *is* highly selective at ω_p, it will suppress undesired harmonics of the pump voltage that may be generated in the process of pumping the nonlinear diode.

Let the desired input voltage and current be sinusoids having the representation

$$\left.\begin{aligned} v_1(t) &= V_1 \exp(j\omega_1 t) + V_1^* \exp(-j\omega_1 t) \\ i_1(t) &= I_1 \exp(j\omega_1 t) + I_1^* \exp(-j\omega_1 t) \end{aligned}\right\} \quad (1.10)$$

The corresponding expressions for the output voltage and current are

$$\left.\begin{aligned} v_2(t) &= V_2 \exp(j\omega_2 t) + V_2^* \exp(-j\omega_2 t) \\ i_2(t) &= I_2 \exp(j\omega_2 t) + I_2^* \exp(-j\omega_2 t) \end{aligned}\right\} \quad (1.11)$$

The total signal voltage $v(t)$ across the diode is the sum of $v_1(t)$ and $v_2(t)$. The total signal current $i(t)$ is composed of the desired signal components $i_1(t)$ and $i_2(t)$ and a remainder consisting of all the sideband currents [indicated by $i_r(t)$ in Fig. 3].

To be specific we shall assume that the input frequency is above the pump frequency and the output frequency below it, and that mixing occurs via the fundamental or first harmonic of the pump frequency, that is, $\omega_2 = \omega_1 - \omega_p$. In other words, the mixer is non-inverting. This will not affect the generality of the analysis since our final results will apply to all forms of up- and down-conversion.

Inserting equations (1.10) and (1.11) into (1.6), and equating terms of like frequency, we obtain the pair of equilibrium equations

$$\begin{aligned} I_1 &= (g_0 + j\omega_1 C)V_1 + g_1 V_2 \\ I_2 &= g_1 V_1 + (g_0 + j\omega_2 C)V_2 \end{aligned} \quad (1.12)$$

relating the sinusoidal amplitudes. These equations represent the signal conversion properties of the mixer. They can be expressed in matrix form as

$$\begin{bmatrix} I_1 \\ I_2 \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \quad (1.13)$$

where the admittance matrix is given by

$$\begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} = \begin{bmatrix} g_0 + j\omega_1 C & g_1 \\ g_1 & g_0 + j\omega_2 C \end{bmatrix} \quad (1.14)$$

This matrix, which also applies to the non-inverting up-converter, represents the two-port network shown in Fig. 4. The junction capacitance, by the mixing action, shows its effect at both signal frequencies. Notice that the conversion matrix obeys reciprocity, i.e. $y_{12} = y_{21}$. This would not necessarily be the case if a sideband voltage were present, since the conductance coefficients in the resultant 3×3 matrix can not always be made real by a shift in the time axis.

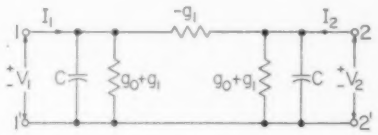


FIG. 4. Small-signal equivalent circuit of a non-inverting mixer.

The conversion matrix for the two types of inverting mixer is

$$\begin{bmatrix} g_0 + j\omega_1 C & g_1 \\ g_1 & g_0 - j\omega_2 C \end{bmatrix} \quad (\omega_1 + \omega_2 = \omega_p)$$

where $\omega_1 > \omega_2$ for the down-converter and $\omega_1 < \omega_2$ for the up-converter. The only difference from the preceding matrix is in the sign preceding ω_2. The applicable circuit variables in this case are V_1, I_1, and V_2^*, I_2^*.

Before proceeding with the derivation of explicit mixer properties, we shall consider two general properties of a two-port network which are pertinent to our analysis. The first concerns the activity of the two-port, that is, its ability to exhibit power gain; the second relates to its stability under prescribed loading conditions at its signal ports.

1.2 *Mixer activity*

It can be shown[9] that a two-port having the admittance matrix

$$\begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix}$$

can exhibit power gain between its two ports if, and only if, at least one of the following three conditions is fulfilled:

$$\mathrm{Re}\, y_{11} = g_{11} < 0 \quad (1.15a)$$

$$\mathrm{Re}\, y_{22} = g_{22} < 0 \quad (1.15b)$$

$$g_{11} g_{22} < \left| \frac{y_{12} + y_{21}}{2} \right|^2 \neq 0 \quad (1.15c)$$

where "Re" denotes the real part.

Applying these criteria to the mixer we obtain

$$g_0 < |g_1| \neq 0 \tag{1.16}$$

as the necessary condition for gain. Thus, gain is always possible if the average value of the time-varying conductance is negative and also if it is positive but less than the magnitude of g_1. These conclusions are sketched in Fig. 5. Using the

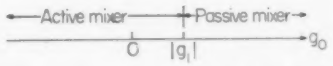


FIG. 5. Pertinent to activity of an Esaki mixer.

integral formula for the Fourier coefficients and the definition of $g(t)$, one can show that the activity condition can only be fulfilled if the slope $\partial I/\partial V$ changes sign or is always negative in the region of operation, i.e. in the region of pump voltage excursion.

Fig. 6 illustrates three possible sinusoidal pumping conditions that lead to a mixer with gain. In the first diagram, Fig. 6(a), the diode is biased in the positive slope region (positive conductance region) so that $g_0 > 0$. If the bias point is moved into the region of negative slope near the point of minimum negative conductance, then $g_0 < 0$, as illustrated in Fig. 6(b). Notice the radical change in the waveform of $g(t)$. In Fig. 6(c) the diode is biased at the boundary between the positive and negative slope regions, i.e. at the point where I is a maximum, or $\partial I/\partial V = 0$. For small pump excursions about this point $g_0 = 0$.

For later reference, notice that since the *slope* of $\partial I/\partial V$ is negative over most of the operational region in the three cases cited, the fundamental component of $g(t)$, i.e. the component at the pump frequency, is 180° out of phase with the pump voltage. In other words g_1 is negative.

If the diode were pumped about its minimum conductance point with a sufficiently small amplitude so that a parabolic approximation of $\partial I/\partial V$ would hold, the lowest frequency component in $g(t)$ would be $2\omega_p$, not ω_p, since $g_1 = 0$. In this case one could use "subharmonic" pumping. More will be said of this in Section 5.2.

1.3 *Mixer stability*

An active two-port, in the general case (i.e. $g_{11}, g_{22}, \lesseqgtr 0, y_{12}, y_{21} \neq 0$), can also oscillate (have infinite gain) unless it is appropriately loaded by passive terminations at its ports. Since only a stable mixer is of significance, we must establish the bounds of allowable terminations. To do this we first define what we mean by a stable (linear) network. A network, including its external passive loading, is stable if, and only if, it will produce a response which does not increase with time when energized by an arbitrary, finite excitation such as internal noise. The borderline between stability and instability corresponds to a response which just fails to grow with time. Alternately, if stated in terms of the frequency domain, a network is stable if the response to a sinusoid of arbitrary frequency and finite amplitude is also finite; or equivalently, the response to a vanishingly small excitation is zero.

It can be shown[9] that in order for a two-port plus its terminations (Fig. 7) to be stable, it is necessary that the following three conditions be satisfied:

$$G_1+g_{11} > 0 \tag{1.17a}$$

$$G_2+g_{22} > 0 \tag{1.17b}$$

$$(G_1+g_{11})(G_2+g_{22}) > \frac{|y_{12}y_{21}| + \mathrm{Re}(y_{12}y_{21})}{2} \tag{1.17c}$$

where conductances G_1 and G_2 are the real (dissipative) parts of the external loads at ports 1 and 2, respectively. These conditions define the allowable range of external loading, i.e. of G_1 and G_2, for stability in the sense defined above; in other words, they insure that the real parts of the total node admittances at either signal port will be positive, i.e. $\mathrm{Re}(Y_{in}+Y_1)$, $\mathrm{Re}(Y_{out}+Y_2)>0$. No restrictions are imposed on the imaginary parts of the loads.*

Applying the stability criteria to the mixer we obtain the conditions

$$G_1+g_0 > 0 \tag{1.18a}$$

* Actually, these stability conditions are not "sharp". A precise criterion of stability must be based on the location of the singularities of the total node admittances in the complex frequency plane. However, this latter test is impractical for most non-trivial two-ports. This is particularly true for mixers, where two linearly dependent frequencies are involved. The conditions derived above, however, should be adequate in most practical cases.

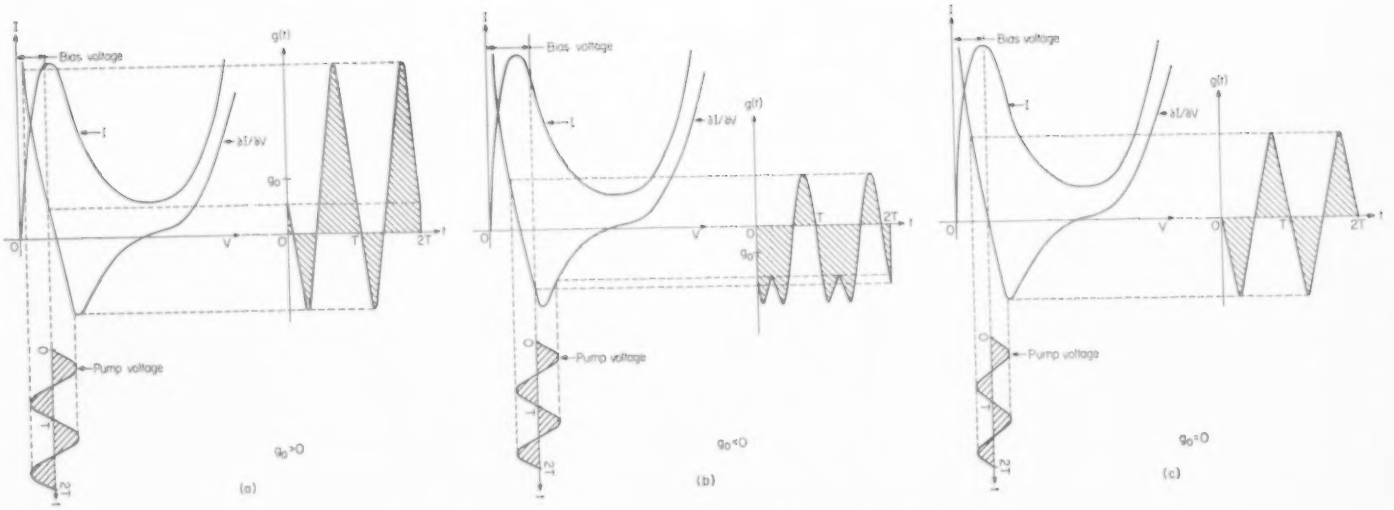
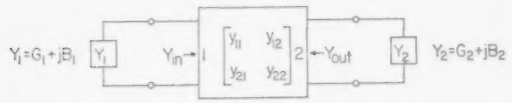


FIG. 6. Possible sinusoidal pumping conditions of an Esaki mixer.

$$G_2+g_0 > 0 \qquad (1.18b)$$

$$(G_1+g_0)(G_2+g_0) > g_1^2 \qquad (1.18c)$$

If any one of these is violated, the mixer will be unstable and will oscillate in the signal circuits.



[FIG. 7. Relevant to the discussion of circuit stability.

Notice, incidentally, that the active mixer is open-circuit unstable ($G_1 = G_2 = 0$), but short-circuit stable ($G_1, G_2 \to \infty$).

2. DERIVATION OF GAIN, GAIN-BANDWIDTH PRODUCT, AND OTHER MIXER PROPERTIES

For concreteness we shall consider a non-inverting down-converter $\omega_1 = \omega_2+\omega_p$, though our final results are independent of this assumption. We postulate that the mixer is terminated in a simple resonant circuit at either port as indicated in Fig. 8. The inductors L_{11} and L_{22} are chosen to resonate the port capacitors C_1+C and C_2+C at the center frequencies $\omega_1 = \Omega_1$ and $\omega_2 = \Omega_2$, respectively, where $\Omega_1 = \Omega_2+\omega_p$. The capacitors C_1 and C_2 represent any externally applied and stray capacitances. The signal source, assumed to be sinusoidal, is represented by the current source I_s and the conductor g_s. The conductors g_l, and g' and g'' correspond to the load and circuit losses, respectively.

If we denote the total node admittance at either port by the symbols Y_{11} and Y_{22}, then

$$Y_{11} = G_{11}+j\omega_1 C_{11}+(j\omega_1 L_{11})^{-1} \qquad (2.1a)$$

$$Y_{22} = G_{22}+j\omega_2 C_{22}+(j\omega_2 L_{22})^{-1} \qquad (2.1b)$$

where

$$G_{11} = g_s+g'+g_0$$

$$G_{22} = g_l+g''+g_0$$

$$C_{11} = C+C_1$$

$$C_{22} = C+C_2$$

2.1 *Stability*

The mixer will be stable provided that the source and load conductances g_s and g_l satisfy the conditions

$$g_s+g'+g_0 > 0 \qquad (2.2a)$$

$$g_l+g''+g_0 > 0 \qquad (2.2b)$$

$$(g_s+g'+g_0)(g_l+g''+g_0) > g_1^2 \qquad (2.2c)$$

The domain of stability is indicated in the g_s-g_l plane in Fig. 9 for the two cases $g_0 \lesseqgtr 0$ and the low-loss situation $g', g'' \simeq 0$. As is evident from Fig. 9(a), the source and load conductances must exceed $|g_0|$ for the case $g_0 < 0$, to insure stability.

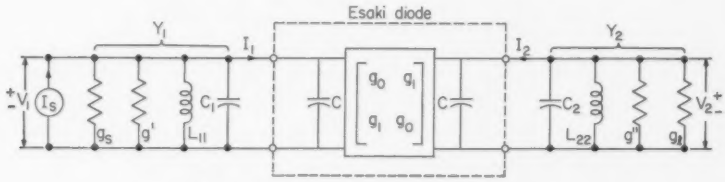
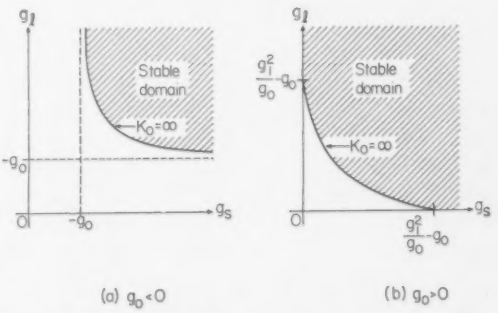


FIG. 8. A tuned non-inverting mixer.



(a) $g_0<0$ (b) $g_0>0$

FIG. 9. The permissible domain of source and load conductances for stable operation.

In either case, the boundary line separating the stable and unstable domains corresponds to an infinite conversion gain. This boundary is defined

by equation (2.2c) when the inequality sign is replaced by an equality sign.

2.2 *Gain*

Of the many possible ways to specify the gain of a two-port, we choose for convenience "transducer gain". The transducer gain K is defined as the average power actually delivered to the load g_l, divided by the maximum average power that can be absorbed from the source. This latter power, called the available power of the source, is given by $|I_s|^2/4g_s$ where $|I_s|$ is the r.m.s. value of the sinusoidal current source. Letting $|V_2|$ denote the r.m.s. value of the signal voltage developed across the load conductance, we obtain for the conversion gain the expression

$$K = \frac{g_l|V_2|^2}{|I_s|^2/4g_s} = 4g_s g_l\left|\frac{V_2}{I_s}\right|^2 \tag{2.3}$$

The ratio $|V_2/I_s|$ can be obtained from the solution of the equilibrium equations (1.12), and the terminating conditions

$$\left.\begin{aligned} I_s &= I_1 + Y_1 V_1 \\ I_2 &= -Y_2 V_2 \end{aligned}\right\} \tag{2.4}$$

(See Fig. 8.) Thus,

$$K = \frac{4g_s g_l g_1^2}{|Y_{11}Y_{22} - g_1^2|^2} \tag{2.5}$$

This expression permits us to evaluate the conversion gain as a function of frequency.

If we denote the frequency deviation from resonance by $\Delta\omega$, that is

$$\Delta\omega = \omega_1 - \Omega_1 = \omega_2 - \Omega_2 \tag{2.6}$$

we may express Y_{11}, Y_{22} and K as functions of $\Delta\omega$. The resultant expressions are quite unwieldy but may be simplified if the fractional deviations, $\Delta\omega/\Omega_1$ and $\Delta\omega/\Omega_2$ over the operating bandwidth of the mixer are small in magnitude compared to unity. Since $\Omega_1 > \Omega_2$, these conditions will be satisfied if the fractional operating bandwidth* at the output port is much less than unity. This is usually true for practical down-converters. Assuming this to be the case, then Y_{11} and Y_{22} can be represented adequately by the simplified formulas

$$\left.\begin{aligned} Y_{11} &= G_{11} + j2\Delta\omega C_{11} \\ Y_{22} &= G_{22} + j2\Delta\omega C_{22} \end{aligned}\right\} \tag{2.7}$$

The expression for conversion gain takes the form

$$K(x) = \frac{4g_s g_l g_1^2}{a^2 + bx^2 + x^4} \tag{2.8}$$

where x is a frequency variable given by

$$x = 2\Delta\omega\sqrt{(C_{11}C_{22})}$$

and where

$$a = G_{11}G_{22} - g_1^2 = (g_s + g' + g_0)(g_l + g'' + g_0) - g_1^2$$

$$b = \left[G_{22}\sqrt{\left(\frac{C_{11}}{C_{22}}\right)} + G_{11}\sqrt{\left(\frac{C_{22}}{C_{11}}\right)}\right]^2 - 2a$$

The gain at midband, $x = 0$, is given by

$$K(0) \equiv K_0 = \frac{4g_s g_l g_1^2}{a^2} \tag{2.9}$$

It is evident that in order for the gain to be high, the denominator must be very small, i.e. $a \simeq 0$. If losses are low enough, this can be achieved by a proper choice of g_l and g_s since $g_0 < |g_1|$. In the limit $a = 0$, the gain becomes infinite and the converter will oscillate simultaneously at Ω_1 and Ω_2. This limit corresponds to the boundary of the stable domain in Fig. 9.

2.3 *Gain-bandwidth product*

The expression for the bandwidth B, the frequency interval about resonance within which the conversion gain is equal to or greater than half its midband value, is derived by solving for the two real values of x for which the denominator in equation (2.8) is twice its midband value. The two solutions are equal and opposite in sign. Using these solutions and the definition of x, we obtain the formula

$$B = \frac{1}{2\pi\sqrt{(C_{11}C_{22})}}\,\frac{a}{\{\sqrt{[(b/2)^2 + a^2]} + b/2\}^{1/2}} \tag{2.10}$$

Notice that the bandwidth vanishes in the limit of high gain $a \to 0$. This is to be expected since the converter is a negative conductance device.

* The operating bandwidth will usually be less than the "cold" bandwidth of the tuned circuit because of the negative output conductance of the mixer.

Since the bandwidth is gain dependent, it is desirable to express it explicitly in terms of the midband gain. This can be done (after considerable computational effort), but the resulting expression is very cumbersome. However, in the limit of high gain $a \simeq 0$ one obtains a rather simple and useful expression. To derive it we must expand equation (2.10) in a power series in a, bearing in mind that b is also a function of a. Carrying out this expansion, we obtain

$$B = \frac{a}{2\pi(G_{11}C_{22}+G_{22}C_{11})} + \dots \quad (2.11)$$

Thus, the bandwidth at high gain decreases inversely as the square root of the midband gain. It follows that the gain-bandwidth product at high gain

$$\sqrt{(K_0)}B = \frac{\sqrt{(g_s g_l)}|g_1|}{\pi(G_{11}C_{22}+G_{22}C_{11})} \qquad (g_0 < |g_1| \neq 0) \quad (2.12)$$

is independent of either gain or bandwidth and hence is a useful figure of merit of the converter. The higher this product, the better the converter.

When losses are negligible, this expression simplifies to

$$\sqrt{(K_0)}B = \frac{\sqrt{(g_s g_l)}\,|g_1|}{\pi[(g_s+g_0)C_{22}+(g_l+g_0)C_{11}]} \quad (2.13)$$

Since this is a function of the external loading, g_s and g_l, it is useful to know what distribution of this loading will yield the maximum product. To answer this question we first remind ourselves that the gain-bandwidth product was derived under the high-gain restriction $(g_s+g_0)(g_l+g_0) \simeq g_1^2$; therefore, we must also maximize it under this constraint. Maximizing equation (2.13) with respect to either g_s or g_l leads to a fourth-degree polynomic equation for the optimum source conductance. This necessitates a numerical solution. It is evident from equation (2.13) that the resultant optimum value of g_s and g_l, when normalized to the modulation conductance $|g_1|$, and the corresponding maximum gain-bandwidth product, when normalized to the bandwidth $|g_1|/(2\pi C_{11})$, can only be functions of the output-input capacitance ratio c,

$$c = \frac{C_{22}}{C_{11}}$$

and the modulation coefficient μ defined by

$$\mu = \frac{g_0}{|g_1|}$$

We have performed the optimization for a large number of μ-values covering the active range $\mu < 1$ and for $c = 1$, 10, 100. The results are summarized in Fig. 10. The curves also apply to the case when $C_{11} > C_{22}$ if the changes $C_{11} \rightleftarrows C_{22}$ and $g_s \rightleftarrows g_l$ are made. Notice from Fig. 10(a) that for a given input capacitance C_{11}, the highest maximum gain-bandwidth product will occur when the input and output capacitances are equal. In this case the optimum source and load conductances are also equal. For this "symmetric loading" situation, the optimal conditions can also be stated analytically. Thus, with $C_{11} = C_{22}$,

$$[\sqrt{(K_0)}B]_{\max} = \frac{|g_1|-g_0}{2\pi C_{11}} \quad (2.14a)$$

and

$$g_{s_{\text{opt}}} = g_{l_{\text{opt}}} = |g_1|-g_0 \quad (2.14b)$$

Of course, g_1 is restricted to the active range $g_0 < |g_1| \neq 0$. Naturally if C_{11} and C_{22} have their least possible value, namely the capacitance of the diode alone, then the maximum gain-bandwidth product given above will have its highest possible value,

$$[\sqrt{(K_0)}B]_{\max} = \frac{|g_1|-g_0}{2\pi C} \quad (2.15)$$

Since this quantity is a function only of the diode properties, it is a useful design parameter. Comparing this expression with the corresponding one for the Esaki amplifier,[10] $|g(V)|/2\pi C$, we observe that a diode with a high conductance-to-capacitance ratio is desirable in either application.

As a practical note it is worth mentioning that for the range $\mu < 0$, the numerically computed curves can be represented by the empirical expressions

$$[\sqrt{(K_0)}B]_{\max} = \frac{|g_1|}{\pi C_{11}} \frac{\sqrt{[c^{-0.6}(c^{-0.6}-\mu)(1-\mu c^{-0.6})]}}{1+c^{-0.2}} \quad (2.16a)$$

$$g_{s_{\text{opt}}} = |g_1|(c^{-0.6}-\mu) \quad (2.16b)$$

$$g_{l_{\text{opt}}} = |g_1|(c^{0.6}-\mu) \quad (2.16c)$$

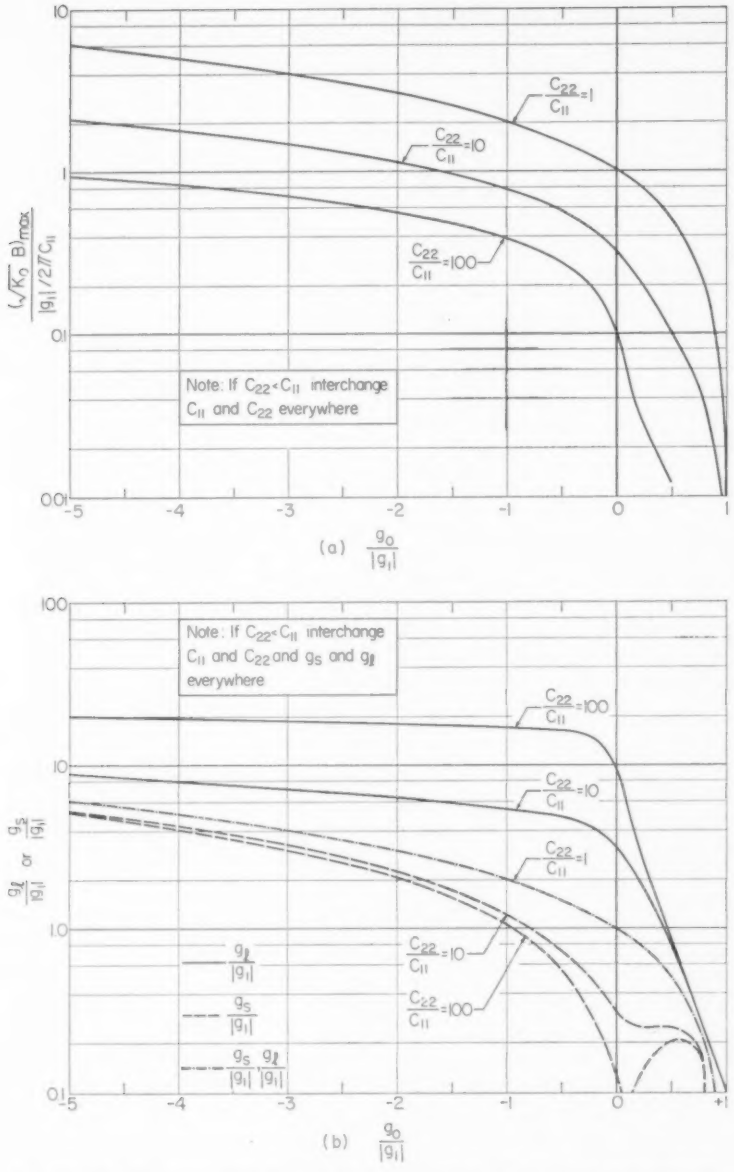


FIG. 10. (a) Maximum gain-bandwidth product at high gain, and (b) optimum source and load conductance as a function of the modulation coefficient and capacitance ratio.

These formulae are known to be accurate to within about ± 10 per cent for the range $-10<\mu<0$, $0{\cdot}01<c\leqslant 100$.

Returning to Fig. 10 and to equation (2.14), we observe that when g_1 is fixed, the gain-bandwidth product increases as the average conductance becomes large and negative. As we shall demonstrate in Section 4 with a numerical example

based on an actual diode characteristic, this means that the highest maximum gain-bandwidth product will be obtained when the Esaki diode is biased in its negative conductance region.

Several important facts should be kept in mind concerning Fig. 10 and the preceding expressions. First, since mixing at the fundamental pump frequency ceases when $g_1 = 0$, the gain-bandwidth product must vanish for this condition. Second, because negligible circuit losses were assumed, the curves and expressions will be in error when losses are present. However, this error should be small, if the losses are small, except possibly near the passive $|g_1| \simeq g_0$ or near the mixing limit $g_1 \simeq 0$. When losses are present the high-gain assumption can still be satisfied near the passive limit, but now only to values of $|g_1|$ somewhat in excess of g_0, the exact amount depending on the degree of losses. For example, for the unlikely case of equal port losses, $g' = g''$, the preceding expressions may again be used provided that the change $g_0 \to g_0 + g'$ is made and the restriction $g_0 + g' < |g_1| \neq 0$ imposed.

The remaining question centers about the behavior of the maximum gain-bandwidth product when the high-gain assumption is not fulfilled. As implied earlier, it is a very tedious numerical calculation to analyze the general case of finite gain and arbitrary g_0. An exception is the special case $g_0 = 0$ for which an exact analytic solution can be obtained rather simply when losses are negligible. The results of this calculation show that the maximum gain-bandwidth product increases as the gain decreases to unity. However, when the gain exceeds 10 dB this product is within approximately 12 per cent of its high-gain asymptote. Since it is probably safe to assume the same trend when $g_0 \neq 0$, we may consider the term "high gain", as applied to our present discussion, to mean gain in excess of 10 dB.

2.4 *Input and output admittance at resonance*

For various practical reasons it is useful to know over what range of values the input and output admittances presented by the mixer at resonance may vary as the gain assumes any possible value. These admittances are purely real at resonance and are given by the expressions

$$Y_{in} = G_{in} = g_0 - \frac{g_1^2}{g_l + g'' + g_0} \qquad (2.17a)$$

$$Y_{out} = G_{out} = g_0 - \frac{g_1^2}{g_s + g' + g_0} \qquad (2.17b)$$

In order to study the behavior of these conductances as functions of the loading g_s and g_l, we first note that since the denominators in the second terms must never become negative for reasons of stability, the maximum value which the input and output conductance may attain is simply g_0, the average value of the time-varying conductance. This means that when $g_0 > 0$, G_{in} and G_{out} can be positive if g_l and g_s are large enough; however, when $g_0 < 0$, G_{in} and G_{out} can never be positive. On the other hand, provided that the losses are not too large, the lowest value of G_{in} and G_{out} is negative irrespective of the sign of g_0, provided that $g_0 < |g_1|$, the condition for an active mixer. The lower bound is a finite negative number when $g_0 > 0$, but is infinite when $g_0 < 0$ if losses are small enough.

It is more instructive to study the behavior of G_{in} and G_{out} not as functions of g_s and g_l, but as functions of each other and of the midband gain K_0. To do this one needs merely to solve equation (2.17) for g_s and g_l and to insert the resulting expressions into the midband gain formula, equation (2.9). We have done this for the low-loss case $g' \simeq g'' = 0$, and have computed a series of constant-gain contours in the $G_{in} - G_{out}$ plane. Before presenting these results it is instructive to note that the range of allowable values of G_{in} and G_{out} in the G_{in}–G_{out} plane is prescribed by the stability conditions and is as indicated by the cross-hatched regions in Fig. 11. The boundary of the region is defined by a segment of the hyperbola $(G_{in} - g_0)(G_{out} - g_0) = g_1^2$, which is the condition for infinite gain, and by a series of straight-line segments which correspond to zero gain. As we explained earlier, and as is evident in Fig. 11, G_{in} and G_{out} can be positive if $g_0 > 0$, but must be negative if $g_0 < 0$. Notice the finite lower bound on G_{in} and G_{out} in the former case.

For purposes of illustration, a series of constant-gain contours have been computed for several positive and negative values of μ (or g_0) and the results are illustrated in Fig. 12. Since losses were neglected, all contours exhibit symmetry about the 45°-line, i.e. about the line $G_{in} = G_{out}$. Observe that for $\mu > 0$, G_{out} exhibits a maximum

positive value when G_{in} is negative. This positive value is given by

$$\left(\frac{G_{out}}{|g_1|}\right)_{max} = \mu + \frac{K_0}{\mu}\left\{\sqrt{\left[\left(\frac{1+K_0^{-1}}{2}\right)^2 - \frac{\mu^2}{K_0}\right]} - \frac{1+K_0^{-1}}{2}\right\} \tag{2.18}$$

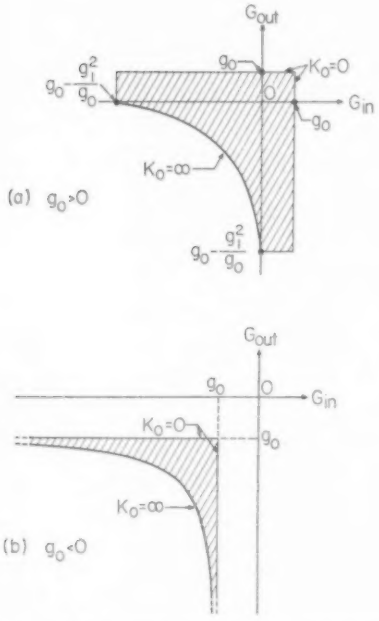


FIG. 11. Range of input and output conductance at resonance.

This cannot exceed $0{\cdot}5\,g_0$, if $K_0 > 1$. It is a curious fact that the corresponding value of g_l is also given by this equation. This suggests that the mixer is "matched" at the output, but, of course, this conclusion is meaningless here, since matching has no significance for a negative conductance device. The above expression also holds for the maximum positive input conductance.

The conditions along the symmetry line are easily obtained. They are

$$\frac{G_{in}}{|g_1|} = \frac{G_{out}}{|g_1|} = \mu - \frac{1}{(g_s/|g_1|)+\mu} \tag{2.19a}$$

where

$$\frac{g_s}{|g_1|} = \frac{g_l}{|g_1|} = \left(\frac{1}{\sqrt{(K_0)}} - \mu\right) + \sqrt{\left[\left(\frac{1}{\sqrt{(K_0)}} - \mu\right)^2 + 1 - \mu^2\right]} \tag{2.19b}$$

Clearly G_{in}, G_{out}, g_s and g_l are all decreasing functions of gain. With the help of this equation it is easily proven that when $g_0 < 0{\cdot}5|g_1|$, the mixer cannot *simultaneously* present a positive input and a positive output conductance at resonance and at the same time exhibit gain. More precisely, the midband gain must be less than $4\mu^2$ when the input and output conductances are positive, simultaneously. Thus, since $\mu < 1$ for an active mixer, the maximum obtainable midband gain for this condition is less than 6 db. These conclusions are altered somewhat if losses are present.

3. MIXER NOISE THEORY

The ultimate sensitivity of a microwave receiving system is to a large extent determined by the amount of noise generated within the mixing element and its associated circuitry. Consequently, on practical grounds a study of mixer noise performance is of prime importance.

The noise associated with the mixer circuitry is usually generated within ohmic elements and therefore is thermal in nature and consequently simple to evaluate. Often the noise attributed to the vacuum tube which may follow the mixing element is also included as part of the mixer noise. Though non-thermal in origin, this contribution is easily calculated. The noise generated within the mixer element, in this case the Esaki diode, is of three forms: thermal, $1/f$ and shot. The thermal noise is produced within the diode series resistance R_b and can be minimized by proper diode design and by operating the diode sufficiently below its resistive cutoff frequency.[10] The $1/f$ noise is believed to be associated with tunneling via intermediate states in the band gap and hence is of bulk origin. According to the meager experimental results reported to date, this noise is most predominant at the lower end of the negative conductance region and in the broad valley region beyond,[11] and thus can be

minimized by restricting the pump voltage excursions from entering this region. The shot noise is by far the most important source of noise in a well-designed and properly operated Esaki mixer diode. The level of this noise, which originates in the random tunneling of electrons across the *p–n* junction, is determined by the magnitude of the tunnel currents composing the net diode current. Under pumped conditions, certain correlation effects arise in the mixer which may lead to partial shot-noise cancellation. To explain these effects properly, and to assess correctly the dominant role played by the shot noise, it is necessary to discuss its origin in some detail.

3.1 *Theory of internal mixer noise*

The shot noise generated in an Esaki diode consists of small random fluctuations superimposed on the macroscopically observed diode tunnel current. By macroscopic current we mean the current registered on some averaging device, such as an ammeter. The usual formula obtained for the mean-square value of this noise is predicated on the assumption that the statistics governing the noise fluctuations are independent of time or nearly so; that is, the random process is considered stationary. This would be the case, for example, if the tunnel current were constant with time, or only weakly varying with time as under small-signal amplifier conditions. However, this is not

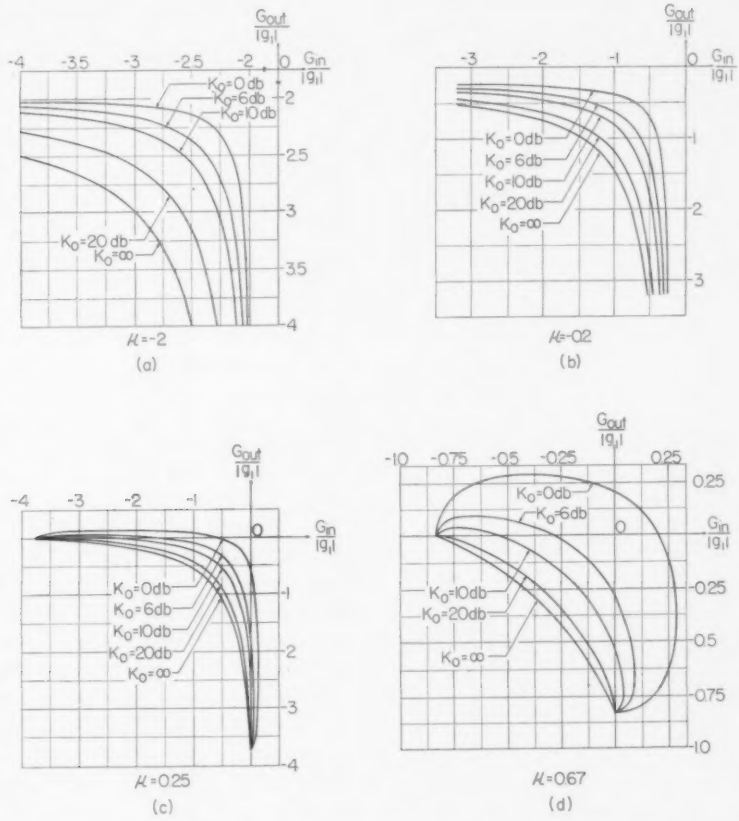


FIG. 12. Constant-gain contours in the G_{in}–G_{out} plane for various values of the modulation coefficient.

true under mixing conditions since the tunnel current varies strongly about some average value. In this case, we have a nonstationary random process and we cannot extend the stationary results to it by simply letting the noise spectrum become time-dependent, since this is meaningless from both the theoretical and physical standpoint.

The observed tunnel current in an Esaki diode, in actuality, consists of two oppositely directed components as indicated in Fig. 13(a). One component I_e is caused by electrons which tunnel across the junction from the n- to the p-side.

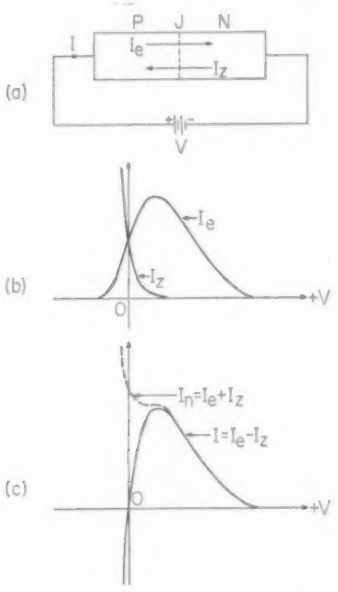


FIG. 13. Definitions and sketches of the tunnel currents of an idealized Esaki diode.

This component, which is responsible for the negative conductance, is illustrated in Fig. 13(b) for an idealized Esaki diode. The oppositely directed component I_z, also illustrated in this figure, is responsible for the well-known Zener "breakdown". The net diode current $I(V)$ is equal to the difference of these components

$$I = I_e - I_z \tag{3.1}$$

The quantity of interest in noise computations, we shall see, is not the difference of the tunnel components, but rather the sum. In anticipation of its importance in noise considerations we shall denote this sum current by the subscript n. Hence

$$I_n = I_e + I_z \tag{3.2}$$

This current is sketched in Fig. 13(c) along with the diode current. By its definition I_n is necessarily non-negative. It can be shown[12] that for an idealized Esaki diode, I_n and I are simply related by the expression

$$I_n = I \coth \frac{qV}{2kT} \tag{3.3}$$

where q is the electronic charge ($q = 1{\cdot}6 \times 10^{-19}$ C), V the applied diode voltage, T the absolute temperature of the diode and k Boltzmann's constant ($k = 1{\cdot}38 \times 10^{-23}$ J/deg). This expression is believed to be valid for practical diodes, also.

Either tunnel component consists of the combined contribution of all electrons which cross the junction in one of the two directions. Whenever an electron crosses the junction, it produces an elementary current pulse whose duration is determined by the time taken to cross the junction and whose algebraic sign is governed by the direction of crossing. The only difference between elementary pulses, aside from their algebraic sign, is their time of occurrence, which is a "pure" random variable.* In other words, either tunnel current is a sum of identical elementary pulses distributed in time in a purely random fashion. A typical elementary pulse, $j(t-t_k)$, is zero before its respective emission time $t = t_k$, and corresponds to one electronic charge; that is, it satisfies the conditions

$$\left.\begin{aligned} j(t-t_k) &\geqslant 0 \quad t \geqslant t_k \\ &= 0 \quad t < t_k \end{aligned}\right\} \tag{3.4a}$$

$$\int_{-\infty}^{\infty} j(t-t_k)\,dt = q \tag{3.4b}$$

Formally, we may decompose each tunnel component into a noise-free or "macroscopically

* By "pure" random variable we mean that the probability of an electron tunneling at a given time is not a function of the number which have tunneled previously or the times at which they have tunneled. This excludes any type of noise "smoothing" mechanism. See Ref. 13 for details.

smooth" part, $I_e(t)$ or $I_z(t)$, plus a "noisy" or fluctuating part. Our objective is to show that certain mean-square values of the "noisy" components are expressible entirely in terms of the noise-free parts. The approach we use, though similar in some respects to the development presented by UHLIR[14], differs considerably from the methods employed by others[15,16].

Suppose we consider the waveform of one of the tunnel components during a small but arbitrary time interval $d\theta$ extending from $t = \theta$ to $t = \theta + d\theta$. This interval is assumed to be small compared to the duration of a single elementary pulse but large enough to contain many emission times. Imagine that one observes, for the same time interval, waveforms of a large group or ensemble of "identical" diodes, which in the limit are infinite in number. Though one would find that the number of emissions during the given interval would differ from waveform to waveform, nevertheless, this set of numbers would cluster about an average value. This average value, known as the statistical or ensemble average, is in general a function of the particular time of observation θ. We shall therefore denote it as $dN(\theta)$. If the observation interval is short enough, the average number of emissions during that interval will be proportional to the length of the interval because of the statistical laws (Poisson statistics) obeyed by the shot process.[17] Since the factor of proportionality is a function of time, we shall designate it as $\rho(\theta)$. Thus

$$dN(\theta) = \rho(\theta)d\theta$$

It is quite evident that $\rho(\theta)$ is simply the statistically averaged rate of emission at the time $t = \theta$. With this interpretation in mind, it follows that the statistically averaged tunnel component, that is, the fluctuation-free part, can be represented at time t by a superposition integral of the form[17,18]

$$\int_{-\infty}^{t} \rho(\theta)j(t-\theta)d\theta$$

where the upper limit can be made infinite because of equation (3.4a).

Physically interpreted, this integral states that electrons that tunnel during the interval θ, $\theta+d\theta$, at the rate $\rho(\theta)$ per second, and number $\rho(\theta)d\theta$, produce a current $[\rho(\theta)d\theta]\ j(t-\theta)$ at the later time t. Integration over θ simply sums the effects of all electrons that have tunneled prior to the time t. Using this result and denoting the emission rates of the two tunnel components by $\rho_e(t)$ and $\rho_z(t)$, we may express the sum current, equation (3.2), by the integral

$$I_n(t) = \int_{-\infty}^{\infty} \rho_n(\theta)j(t-\theta)d\theta \tag{3.5}$$

where $\rho_n(\theta)$ is the sum of the two emission rates.

Since the time taken for tunneling (if one may even speak of a tunneling time) is negligible compared to the period of any relevant signal or pump variation, we may represent each elementary pulse by a unit delta function $\delta(t-t_k)$* scaled by the electronic charge, that is

$$j(t-t_k) = q\delta(t-t_k)$$

Therefore, ρ_n and I_n are simply related as

$$\rho_n(t) = q^{-1}I_n(t) \tag{3.6}$$

Turning attention to the mixer noise associated with the current fluctuations, we first note that each elementary pulse necessarily contains frequency components over the entire frequency band. The frequency spectrum of $j(t-t_k)$ is given by the Fourier transform,

$$s(\omega,t_k) = \int_{-\infty}^{\infty} j(t-t_k)\exp(-j\omega t)\,dt = q\exp(-j\omega t_k) \tag{3.7}$$

This flat or "white" noise frequency distribution is sketched in Fig. 14(a) for the positive frequency range. Only certain portions of this spectrum are of importance in the mixer as will be shown below.

Imagine that we decompose the above elementary pulse into a sum of three sub-elementary

* The unit delta function is defined by its properties which are

(a) $$\left.\begin{aligned}\delta(t-x) &\neq 0 \quad t = x\\ &= 0 \quad \text{elsewhere}\end{aligned}\right\}$$

and

(b) $$\int_{-\infty}^{\infty} \delta(t-x)f(t)dt = f(x)$$

pulses j_1, j_2 and j_r, in such a manner that j_1 corresponds to all of the frequency components contained in a band centered about the input resonance frequency Ω_1, i.e. in the input pass band. This spectral band is indicated by the cross-hatched segment labeled $s_1(\omega, t_k)$ in Fig. 14(a). Similarly, j_2 corresponds to the band $s_2(\omega, t_k)$ centered about the output resonance frequency Ω_2. Thus j_r represents the remaining portions of the spectrum. Because of the filtering, j_1 and j_2 have a zero average value, that is, they "carry" no net charge.

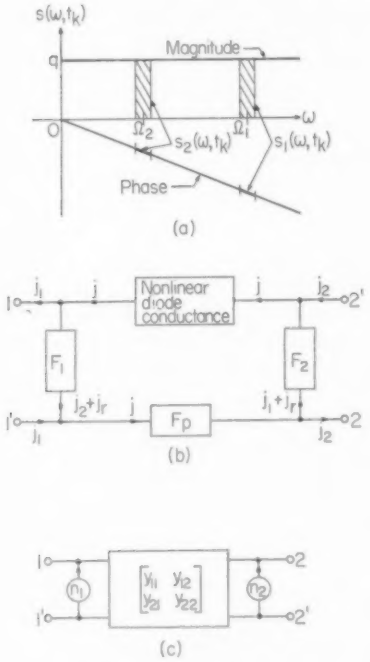


FIG. 14. (a) Frequency spectrum of an elementary current pulse, (b) model used for computation of equivalent noise generators, and (c) circuit representation of equivalent noise generators.

As these three pulses pass through the diode circuitry [Fig. 14(b)], j_2 and j_r will be bypassed through the shorting filter F_1 at the input terminals. On the other hand, j_1 is blocked by this filter and must pass through the input terminals 1–1′. In doing this, it will develop a voltage pulse across the external input circuitry. The spectrum of this voltage pulse will also be centered about Ω_1. Similarly, j_r and j_1 are bypassed by the output filter F_2, and only j_2 will pass through the output terminals 2–2′ to develop a voltage pulse having a spectrum centered at Ω_2. These two voltage pulses then interact with the time-varying conductance as if they were externally applied signals.

What is true for a single elementary pulse is evidently true for a random distribution of such pulses. Thus one may formally decompose the total diode noise current into three parts n_1, n_2 and n_r which obey the same spectral conditions as the corresponding j's. Therefore, as far as the mixing action is concerned, the effect of the diode shot noise is precisely the same as that produced by two noise current generators n_1 and n_2 applied to the input and output terminals as indicated in Fig. 14(c). The correctness of this representation can be verified by noting that it leads to the same short-circuit noise currents at the two signal ports as are obtained from the model, Fig. 14(b).

The input-noise generator n_1, through the mixing action, will produce a current in the output band which is proportional to n_1 but translated in frequency. We denote this frequency-translated counterpart by the symbol n_{12}. The latter will produce a load noise power proportional to its square value n_{12}^2 and a cross-power or "beat" term proportional to the product $n_{12}n_2$. This latter quantity is a correlation term. In addition, the output generator will contribute a term proportional to n_2^2. For noise-power computations, we shall actually want the mean values of these quantities suitably averaged over the statistics and over time.

If we assume that the noise fluctuations of the two oppositely directed tunnel currents are statistically independent, then the mean-square value of the output generator, for example, is simply the *sum* of the squared values of the individual elementary output pulses $j_2(t-t_k)$, appropriately weighted by the time-dependent emission rate ρ_n. Since this sum, a statistical average, is a function of time one must integrate over the latter to obtain the mean-square value of n_2.

If we denote the statistical average by a horizontal bar, and the time average by the symbol $\langle\ \rangle$, then we obtain for the mean-square values, the

expressions[17,18]

$$\overline{\langle n_2^2 \rangle} = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} dt \int_{-\infty}^{\infty} \rho_n(\theta) j_2^2(t-\theta) d\theta \qquad (3.8a)$$

Analogously,

$$\overline{\langle n_{12}^2 \rangle} = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} dt \int_{-\infty}^{\infty} \rho_n(\theta) j_{12}^2(t,\theta) d\theta \qquad (3.8b)$$

$$\overline{\langle n_{12} n_2 \rangle} = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} dt \int_{-\infty}^{\infty} \rho_n(\theta) j_{12}(t,\theta) j_2(t-\theta) d\theta \qquad (3.8c)$$

Here $j_{12}(t, \theta)$ represents the elementary input pulse $j_1(t-\theta)$, translated down in frequency by the amount ω_p to the output pass band. That is, $j_{12}(t, \theta)$ corresponds to the spectral distribution

$$s_1(\omega+\omega_p,\theta) = s_1(\omega,\theta) \exp(-j\omega_p\theta) \qquad (3.9)$$

(for the positive frequency range). This is obtained from the spectrum of the input pulse $j_1(t-\theta)$ by the transformation $\omega \to \omega+\omega_p$.

Under operating conditions $I_n(t)$ will vary periodically in accordance with the pump voltage and hence can be expanded in the Fourier series

$$I_n(t) = \sum_{k=-\infty}^{\infty} I_{nk} \exp(jk\omega_p t) \qquad (3.10a)$$

where

$$I_{nk} = \frac{1}{\tau} \int_0^{\tau} I_n(t) \exp(-jk\omega_p t) dt \qquad (3.10b)$$

Using these expressions and the relation between ρ_n and I_n, equation (3.6), one can show, by a straightforward (but lengthy) calculation, that the mean noise "powers", equation (3.8), are equal to

$$\left.\begin{aligned} \overline{\langle n_2^2 \rangle} &= \overline{\langle n_{12}^2 \rangle} = 2qI_{n0}\,\delta f \\ \overline{\langle n_{12} n_2 \rangle} &= 2q\mathrm{Re}I_{n1}\,\delta f \end{aligned}\right\} \qquad (3.11)^*$$

Here δf is a small frequency interval about the input and output resonance frequencies Ω_1 and Ω_2. The self-power terms are proportional to the average sum current—a result identical to that obtained for the stationary case. The correlation term, on the other hand, is proportional to the amplitude of the first harmonic of I_n. This result is characteristic of the nonstationarity of the random process. Its origin can be traced to the additional phase factor $\exp(-j\omega_p\theta)$ in the translated frequency spectrum, equation (3.9). If mixing had occurred via the Nth pump harmonic, the correlation term would be proportional to the Nth harmonic current amplitude I_{nN}.

The degree of correlation is often expressed in terms of the correlation coefficient γ defined by

$$\gamma = \frac{\overline{\langle n_{12} n_2 \rangle}}{\sqrt{(\overline{\langle n_{12}^2 \rangle}\,\overline{\langle n_2^2 \rangle})}} \qquad (3.12a)$$

$$= \frac{\mathrm{Re}I_{n1}}{I_{n0}} \qquad (3.12b)$$

$$= \frac{\int_0^{\tau} I_n(t) \cos\omega_p t\, dt}{\int_0^{\tau} I_n(t)\, dt} \qquad (3.12c)$$

Since $I_n(t)$ is non-negative, γ is restricted to the range $-1 \leqslant \gamma \leqslant 1$. The closer $|\gamma|$ is to unity, the stronger the correlation. A unity correlation coefficient is approached only when the sum current consists of a sequence of periodic pulses which are short in duration compared to the pump period and which coincide with either the peaks of the fundamental pump wave, $\cos\omega_p t \simeq +1$, $(\gamma \to +1)$ or the troughs, $\cos\omega_p t \simeq -1$, $(\gamma \to -1)$.

One may approach a unity coefficient in several ways. One way is to drive the diode "hard" with a large sinusoidal voltage in order that the pump excursion will reach into the highly conducting Zener "breakdown" region. Pronounced conduction must occur only during a small time interval about the negative peaks of the drive. Another way is to drive the diode with a pulse-like pump voltage, that is, a drive rich in pump harmonics. The resultant current sum will be pulse-like also, because I_n is, to a rough approximation, linearly dependent on bias voltage. (See Fig. 13.) In this case the bias should be at the lower end of the negative conductance region and the pump amplitude should be modest-to-large.

* These results can be generalized to the case when the pulse spectrum is not flat.

Because of the approximate linear dependence of I_n on bias, $I_n(t)$ will be nearly sinusoidal if the pump voltage is sinusoidal. If $I_n(t)$ were truly sinusoidal, $|\gamma|$ could not exceed 0·5. In practice, as we shall demonstrate in Section 4, $|\gamma|$ is somewhat less than this for sinusoidal voltage pumping.

3.2. *Mixer noise figure*

The noise performance of a two-port, such as the mixer, is often characterized by the noise figure. Noise figure is a measure of the noise introduced into the input signal by the noisy two-port and its associated circuitry. The lower the noise figure, the better the two-port. Quantitatively the noise figure, denoted by the symbol F, is equal to the total noise power delivered to the load g_l by all noise sources, divided by that portion of this noise which is attributable to the source itself. By convention, the latter is taken as the (unavoidable) thermal noise generated in the source conductance g_s, assumed to be at room temperature $T = T_s$ ($T_s = 300°K$). Thus

$$F = \frac{N_s + N_a}{N_s} = 1 + \frac{N_a}{N_s} \tag{3.13}$$

Here N_s is the (amplified) thermal noise of the source conductance and N_a the noise of all additional sources. The noise figure is usually computed for a narrow frequency band δf over which all circuit parameters can be assumed constant. For tuned two-ports, such as the mixer under consideration, this band is taken at midband, in which case all admittances are real.

Though the noise contributed by the load is usually considered part of the second-stage noise in conventional mixer circuits, it is often included as part of the mixer noise when the mixer is active, as in our case, because of the amplification that this noise receives from the mixer. We shall adopt the latter practice since the calculated load-contributed noise can always be removed later, if desired.

If we consider the mixer load to be simply an ohmic conductance g_l, then all the relevant external noise sources in the mixer are thermal in nature and can be represented as current sources in shunt with the respective conductances as shown in Fig. 15. Keeping in mind that all thermal sources are uncorrelated, we obtain as the resonance noise figure for the circuit shown in Fig. 15 the expression

$$F = 1 + \frac{g_n}{g_s}\left[\left(\frac{g_s + g' + g_0}{g_1} - \gamma\right)^2 + 1 - \gamma^2\right] \tag{3.14}$$

$$+ \frac{g'T'}{g_s T_s} + \frac{g''T'' + g_l T_l}{g_s T_s}\left(\frac{g_s + g' + g_0}{g_1}\right)^2$$

Here

$$g_n = \frac{qI_{n0}}{2kT_s} \tag{3.15}$$

is an equivalent noise conductance. This conductance is a measure of the self-power shot terms. The lower g_n, the lower the shot-noise contribution. Equation (3.14) has been derived by others, also.[7,19]

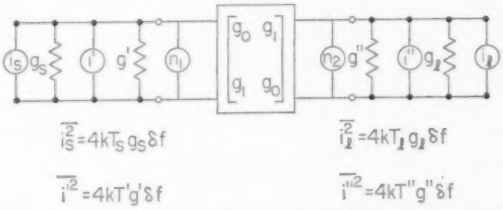


FIG. 15. Circuit used for the calculation of the noise figure at resonance.

The set of terms following unity are the shot-noise contributions. Since $|\gamma| \leqslant 1$, this set is never negative. The remaining terms are thermal contributions. The first is produced by the input loss conductance g' assumed to be at a temperature T'. The next two terms correspond to the output loss conductance g'' at temperature T'' and the load conductance g_l at temperature T_l. If the latter contribution is to be excluded, one may do so by setting T_l to zero.

It is quite evident that partial cancellation of the shot noise can occur if the correlation coefficient has the same sign as g_1. This will usually be true in practice for sinusoidal pumping because g_1 and γ will be negative, since both $I_n(V)$ and $g(V)$ are decreasing functions of V over a large fraction of the usable mixer range. Only when the diode is based near the minimum conductance point will g_1 change sign. In this case some anticompensation will occur.

It is easily shown that when $|\gamma| = 1$, the shot-noise contribution may vanish, depending on the relative sign of γ and g_1. However, in practice, in attempting to approach this ideal condition by the several techniques discussed earlier, one may actually increase the shot-noise contribution since the equivalent noise conductance g_n and the other mixer parameters may be affected adversely. That this can occur will be illustrated in a later Section.

It is not always sufficient to represent the load-contributed noise by a single current generator as we have done. In practice a mixer is usually followed by a vacuum-tube stage (assuming down-conversion). To account properly for all of the tube noise, it is necessary to use two noise generators. One representation, which is particularly convenient for our purposes, consists of a voltage and a current noise generator followed by a noiseless tube as indicated in Fig. 16.[20] The loading effect of the tube at high frequencies caused by transit-time effects is represented by the electronic conductance g_τ in shunt with the current generator. This conductance increases with the square of the frequency.

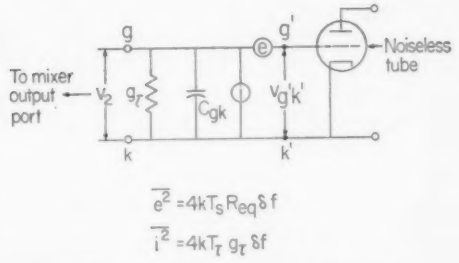


FIG. 16. Circuit representation of the noise properties of a vacuum tube.

The generators represent the shot and partition (of a pentode) noise of the plate circuit and the induced noise of the grid circuit. For computational purposes, it is customary to express these generators in terms of equivalent thermal noise sources. Thus

$$\left.\begin{aligned} \overline{e^2} &\equiv 4kT_s R_{eq}\delta f \\ \overline{i^2} &\equiv 4kT_\tau g_\tau \delta f \end{aligned}\right\} \qquad (3.16)$$

The fictitious resistance R_{eq} is defined, by convention, to be at room temperature. The electronic conductance, on the other hand, is assumed to be at a temperature T_τ sufficiently above room temperature to account for the induced grid noise. In practice $T_\tau \simeq 5T_s$.[20]

To include the tube noise in the noise figure expression, we must compute the additional component of mean-square voltage appearing at the grid-cathode terminals g'–k' of the fictitious noiseless tube, and then divide this result by the corresponding mean-square voltage produced by the source conductance. Since the input admittance of the noiseless tube is zero, we obtain the correction

$$\Delta F = \frac{g_\tau T_\tau}{g_s T_s}\left(\frac{g_s+g'+g_0}{g_1}\right)^2 + \frac{4R_{eq}g_l}{K_0} \qquad (3.17)$$

at resonance. The correlation term between the tube noise generators vanishes near resonance.[20]

If the actual mixer load is assumed, as before, to be a physical (ohmic) conductance g_l external to the tube, then the transit-time conductance simply represents an additional parasitic loss. Therefore, the first correction term can be incorporated with the previous output loss term by letting $g''T'' \to g''T'' + g_\tau T_\tau$ in equation (3.14). Thus only the second contribution represents a distinctly new term. This term will be negligible at very high gain, but may not be at moderate gain (i.e. of the order of 10 dB or less), depending on the value of g_l and R_{eq}. On the other hand, if the transit-time conductance constitutes the actual mixer load, a common condition, then the first term is dropped, and instead the change $g_l T_l \to g_\tau T_\tau$ is made in equation (3.14). Thus, again, only the second contribution represents a true correction term. However, this term now is equal to

$$\frac{4R_{eq}g_\tau}{K_0}$$

since $g_l = g_\tau$. For the commonly used low noise 6AK5 tube, R_{eq} and g_τ have the following values at 30 Mc/s[20]:

$$\begin{aligned} R_{eq} &= 385\ \Omega \text{ (triode)} \\ &= 1880\ \Omega \text{ (pentode)} \\ g_\tau &= 12\ \mu\text{mhos} \end{aligned}$$

Since the product $R_{eq}g_\tau$ is of the order of 10^{-2},

the above term is negligible, even for moderate values of gain. This conclusion also holds if a transformer is used between the mixer and tube to "step up" g_τ to a value more suitable for the diode, since the transformed value of $R_{eq}g_\tau$ does not change. Thus, whether or not the mixer load is considered to be a physical ohmic conductance or the transit-time conductance of the tube, only the second term in equation (3.17) represents a distinctly new contribution to the noise figure expression. This term can always be neglected if the transit-time conductance constitutes the mixer load and if the mixer exhibits gain.

3.3 *Optimization of the noise figure*

In this subsection we shall compute the conditions that must be satisfied by the source and load conductances in order that the noise figure be a minimum. We shall assume that the load is a physical conductance represented by g_l and that all losses are negligible. Under these assumptions

$$F = 1+\frac{g_n}{g_s}\left[\left(\frac{g_s+g_0}{g_1}-\gamma\right)^2+1-\gamma^2\right] +\eta\frac{g_l}{g_s}\left(\frac{g_s+g_0}{g_1}\right)^2 \tag{3.18}$$

where $\eta = T_l/T_s$. That F does exhibit a minimum with respect to g_s follows from the fact that it can be put in the form

$$F = A+Bg_s+\frac{C}{g_s}$$

where B and C are positive. Since the quantities A, B and C are functions of g_l the resulting minimum will also be a function of g_l; therefore, a further minimization with respect to g_l is necessary. However, one is not free to accept the resultant optimum values of g_s and g_l so obtained unless they correspond to a point in the stable domain. (See Fig. 9.) If this point falls in the unstable domain, we shall define the minimum of F to be the lowest value consistent with stability, even though this value does not correspond to an extremum of equation (3.18).

In practice, since g_1 is negative over most of the operating range (at least when sinusoidal pumping is used), we shall let $g_1 = -|g_1|$. However, because of the identity

$$F(-g_1, -\gamma) = F(g_1, \gamma),$$

our results will also apply to positive values of g_1 provided that the change $\gamma\rightarrow-\gamma$ is made.

Of the three groups of terms in equation (3.18), the middle set, representing the internal noise of the diode, is the most interesting from the theoretical viewpoint and the most important from the practical standpoint. A study of the minimal properties of the shot-noise contribution will prove fruitful for several reasons. Firstly, since the lower bound of the total noise figure cannot be less than the lower bound of the shot-noise contribution, the latter bound, when applied to a practical situation, will specify the best noise figure achievable with a given diode operated under a particular set of pumping conditions. Secondly, since the shot-noise contribution is expected to be the dominant one, a knowledge of its properties will provide useful information concerning the properties of the total noise figure. Thirdly, should the output frequency be high enough to permit use of a circulator, such as in double-superheterodyne operation or in up-conversion, the shot noise will represent the entire excess noise of the mixer (aside from incidental noise contributions). Finally, if the load contribution is not to be considered as part of the mixer noise, the shot terms will represent the total excess noise of the mixer.

3.4 *Minimization of the shot-noise figure*

Excluding the load term in equation (3.18), we obtain

$$F_s = 1+\frac{g_n}{g_s}\left[\left(\frac{g_s+g_0}{g_1}\right)^2+2\gamma\frac{g_s+g_0}{|g_1|}+1\right] \tag{3.19}$$

The subscript $_s$ denotes that this expression represents the shot terms only. It is clear that F_s is an increasing function of γ. Since the load conductance does not appear in this expression, the minimization procedure is particularly simple. The only stability condition that might be restrictive is the first one in equation (2.2), namely,

$$g_s+g_0>0$$

Because we are free to choose the value of the load conductance, the remaining two stability

conditions,

$$g_l+g_0>0$$
$$(g_s+g_0)(g_l+g_0)>g_1^2$$

can always be satisfied by selecting g_l large enough. Furthermore, when $g_0>0$, even the first condition is not restrictive, since F_s (and F) always exhibits a minimum for a positive value of g_s.

The minimum of F_s occurs for

$$g_s=|g_1|\sqrt{\left[\left(\frac{g_0}{g_1}\right)^2+2\gamma\frac{g_0}{|g_1|}+1\right]}$$

Obviously, this solution is always valid when $g_0>0$. When $g_0<0$, it is only applicable if it is greater than $-g_0$. This will be the case if the correlation coefficient does not exceed a critical value γ_{0s} given by

$$\gamma_{0s}=-\frac{|g_1|}{2g_0}$$

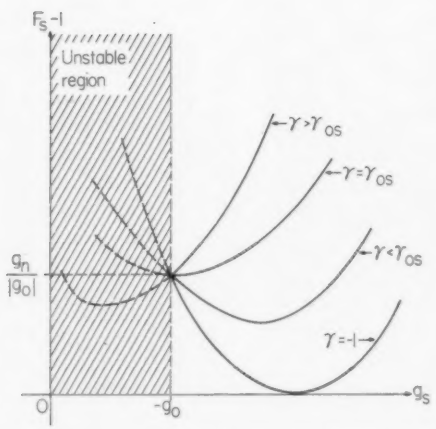


FIG. 17. Excess shot-noise figure as a function of the source conductance and the correlation coefficient for the case $g_0<0$.

Thus, when $g_0>0$ or when $g_0<0$ and $\gamma<\gamma_{0s}$, the minimum value of F_s is calculated by using the optimum value of g_s given above. For the remaining region $g_0<0, \gamma>\gamma_{0s}$, the minimum occurs in the unstable range. This situation is illustrated in Fig. 17. In this case, the lowest value of F_s consistent with stability is realized when g_s exceeds $-g_0$ by an infinitesimal amount. The load conductance necessary for stability becomes arbitrarily large in this limit.

All of the preceding results may be summarized in convenient form in terms of the modulation coefficient

$$\mu=\frac{g_0}{|g_1|}$$

introduced earlier. Denoting the optimum source conductance by $g_{s_{opt}}$ and the corresponding minimum shot-noise figure by $F_{s_{min}}$, we obtain:

for $g_0>0$;

$$\left.\begin{aligned} g_{s_{opt}}&=|g_1|\sqrt{(\mu^2+2\gamma\mu+1)}\\ F_{s_{min}}&=1+\frac{\nu}{\mu}\left[\left(\frac{g_{s_{opt}}}{|g_1|}+\mu\right)^2-1\right]\end{aligned}\right\} \quad (3.20a)$$

for $g_0<0$;

$\gamma\geq\gamma_{0s}$

$$\left.\begin{aligned} g_{s_{opt}}&=|g_0|\\ F_{s_{min}}&=1-\frac{\nu}{\mu}=1+\frac{g_n}{|g_0|}\end{aligned}\right\} \quad (3.20b)$$

$\gamma\leq\gamma_{0s}$

$$\left.\begin{aligned} g_{s_{opt}}&=|g_0|\sqrt{\left(1+\frac{2}{\mu}(\gamma-\gamma_{0s})\right)}\\ F_{s_{min}}&=1+\frac{\nu}{\mu}\left[\left(\frac{g_{s_{opt}}}{|g_1|}+\mu\right)^2-1\right]\end{aligned}\right\} \quad (3.20c)$$

where

$$\gamma_{0s}=-\frac{1}{2\mu};\quad \nu=\frac{g_n}{|g_1|};\quad \mu=\frac{g_0}{|g_1|}$$

If $g_1>0$, let $\gamma\to-\gamma$ everywhere.

With these formulae it is easy to compute the upper and lower bounds of $F_{s_{min}}$. Thus, when $g_0>0$,

$$F_{s_{min}}\leq 1+4\frac{g_n}{|g_1|}(1+\mu) \quad (3.21)$$

This is always less than $1+8g_n/g_0$. On the other hand, when $g_0<0$,

$$F_{s_{min}}\leq 1+\frac{g_n}{|g_0|} \quad (3.22)$$

This latter bound, which corresponds to the limiting value $g_{s_{opt}} = -g_0$, is illustrated in Fig. 17. Jumping ahead, notice that since the load contribution in equation (3.18) vanishes for this value of source conductance, the upper bound expressed by (3.22) is also applicable to the *total* noise figure.

We have plotted the minimum excess shot-noise figure $(F_s-1)_{\min}$ in normalized form in Fig. 18 (solid lines) as a function of γ for a range of modulation coefficients met in practice. From these graphs, it is quite obvious that for a fixed noise conductance and modulation conductance, the best noise figures are obtained when g_0 is negative.

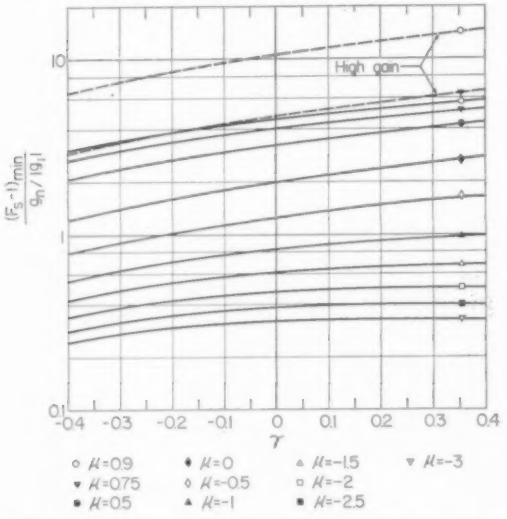


FIG. 18. Minimum excess shot-noise figure as a function of the correlation coefficient and the modulation coefficient.

Arbitrarily high gain is always possible under minimum shot-noise conditions except when the set of conditions, $g_0>0$ and $\gamma>\gamma_{1s}$, is fulfilled, where

$$\gamma_{1s} = \frac{1-3\mu^2}{2\mu^3} \tag{3.23}$$

For this situation, the maximum possible gain is

$$K_{0\max} = \frac{g_1^2 g_s}{(g_s+g_0)[g_0(g_s+g_0)-g_1^2]} \tag{3.24}$$

where g_s has its optimum noise value. This expression is plotted in Fig. 19 as a function of μ for a range of γ-values.

If the mixer gain exceeds the value given above, the lowest shot-noise figure obtainable will exceed the minimum possible value computed earlier.

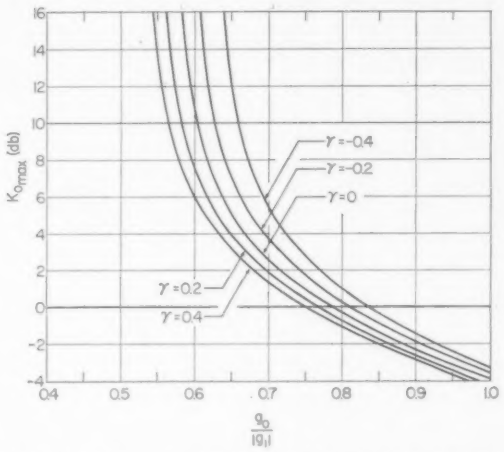


FIG. 19. Maximum midband gain obtainable when the shot-noise figure is minimized.

It is useful to compute the amount of degradation of the minimum when the gain is arbitrarily high. For this case the source conductance must satisfy the high-gain condition

$$(g_s+g_0)(g_1+g_0) \simeq g_1^2 \tag{3.25}$$

This condition restricts the source conductance to the range

$$0 < g_s < g_{s\max} = \frac{g_1^2}{g_0} - g_0 \tag{3.26}$$

[See Fig. 9(b).]

We have indicated this situation in Fig. 20. It is evident from this sketch that the lowest shot-noise figure, consistent with high gain, occurs when g_s equals $g_{s_{max}}$. The corresponding load conductance approaches zero for this case.

In Fig. 18 we have indicated the resulting increase in the minimum shot-noise figure. As is evident, no degradation occurs until μ is sufficiently high to satisfy the inequality $\gamma>\gamma_{1s}$. For the range of γ-values plotted, μ must exceed 0·54 for this to be true. Notice that the amount of degradation increases as the passive limit

$\mu = 1$ is approached. However, the degraded noise figure cannot exceed the value

$$1+\frac{g_n}{g_0}\left(\frac{1+\mu}{1-\mu}\right)$$

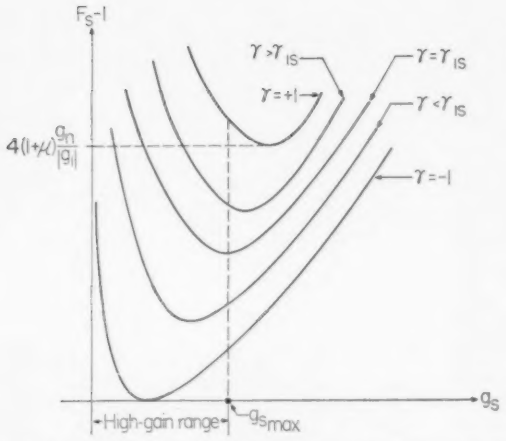


FIG. 20. Excess shot-noise figure as a function of the source conductance and correlation coefficient for the case $g_0 > 0$.

3.5 *Minimization of the total noise figure*

If losses are excluded, the total noise figure is given by

$$F = 1+\frac{g_n}{g_s}\left[\left(\frac{g_s+g_0}{g_1}\right)^2+\gamma\frac{g_s+g_0}{|g_1|}+1\right] + \eta\frac{g_l}{g_s}\left(\frac{g_s+g_0}{g_1}\right)^2 \tag{3.27}$$

Considered as a function of the two independent variables g_s and g_l, F represents a surface. To find the lowest point on this surface, it is necessary to minimize F with respect to g_s and g_l, subject to the condition that these variables never enter the unstable domain indicated in Fig. 9. This two-dimensional minimization coupled with the two-dimensional stability constraints complicates the problem somewhat. However, if the minimization is performed along some judiciously specified path in the g_s-g_l plane, the problem is simplified considerably since it is reduced to essentially a one-dimensional minimization as before. A path having practical importance is the one separating the stable and unstable domains, i.e. the one corresponding to high gain, equation (3.25). Using this relation, we may eliminate g_l in equation (3.27). The resulting extremum of F obtained with respect to g_s represents the minimum total noise figure at high gain.

Keeping in mind the stability condition $g_s > -g_0$ when $g_0 < 0$, and the high-gain threshold condition [equation (3.26)] when $g_0 > 0$, we obtain the following solutions:

for $g_0 > 0$;

$\gamma \geqslant \gamma_{1t}$:

$$\left.\begin{aligned} g_{s_{opt}} &= |g_1|(\mu^{-1}-\mu) \\ F_{min} &= 1+\nu\left(\frac{\mu^2+2\gamma\mu+1}{\mu(1-\mu^2)}\right) \end{aligned}\right\} \tag{3.28a}$$

$\gamma \leqslant \gamma_{1t}$

$$\left.\begin{aligned} g_{s_{opt}} &= |g_1|\sqrt{\left((\mu^{-1}-\mu)^2+\frac{2\nu\mu(\gamma-\gamma_{1t})}{\nu-\eta\mu}\right)} \\ F_{min} &= \left(\frac{\nu}{\mu}-\eta\right)\left[\left(\frac{g_{s_{opt}}}{|g_1|}+\mu\right)^2-1\right]+1-\eta \end{aligned}\right\} \tag{3.28b}$$

for $g_0 < 0$;

$\gamma \geqslant \gamma_{0t}$:

$$\left.\begin{aligned} g_{s_{opt}} &= |g_0| \\ F_{min} &= 1-\frac{\nu}{\mu} = 1+\frac{g_n}{|g_0|} \end{aligned}\right\} \tag{3.28c}$$

$\gamma < \gamma_{0t}$

$$\left.\begin{aligned} g_{s_{opt}} &= |g_0|\sqrt{\left(1+\frac{2\nu(\gamma-\gamma_{0t})}{\mu(\nu-\eta\mu)}\right)} \\ F_{min} &= \left(\frac{\nu}{\mu}-\eta\right)\left[\left(\frac{g_{s_{opt}}}{|g_1|}+\mu\right)^2-1\right]+1-\eta \end{aligned}\right\} \tag{3.28d}$$

where

$$\nu = \frac{g_n}{|g_1|};\quad \mu = \frac{g_0}{|g_1|};\quad \eta = \frac{T_l}{T_s}$$

$$\gamma_{0t} = -\frac{1}{2\mu}\left(1+\frac{\mu\eta}{\nu}\right)$$

$$\gamma_{1t} = \frac{1}{\mu}\left[\left(1-\frac{\mu\eta}{\nu}\right)\left(\frac{1-\mu^2}{2\mu^2}\right)-1\right]$$

If $g_1>0$, let $\gamma\to-\gamma$ everywhere. The corresponding load conductance is obtained from the high-gain condition. The above formulae also apply to the minimum shot-noise figure at high gain if $\eta = 0$.

We have computed $F_{\min}$ as a function of γ for a series of μ-values and a range of $g_n/|g_1|$ ratios. The load was assumed to be at room temperature, $\eta = 1$. The results of this calculation are plotted in Fig. 21. The noise figure is not normalized, since g_n no longer appears as a multiplicative factor in $F_{\min}$. As expected, $F_{\min}$ increases with the noise conductance and also with γ. If g_1 were positive, the opposite dependence on γ would result.

It can be shown that the minimum total noise figure obtained at high gain is always less than the minimum obtained at any lower gain, except when $g_0>0$, $\gamma>\gamma_{1s}$, where γ_{1s} is given by equation (3.23). For this case, the lowest total noise figure is equal to the minimum shot-noise figure. Unfortunately, since the corresponding optimum load conductance vanishes, so does the gain.

In summary, we emphasize that the preceding noise calculations were based on the assumption that all losses were zero and that the mixer load noise could be represented by a simple current generator in shunt with the load. The first assumption is probably not too serious for a well-designed mixer except at frequencies near the resistive cut-off, where thermal noise of the series resistance becomes important. The second may or may not be, depending on the value of the gain, as discussed in Subsection 3.2.

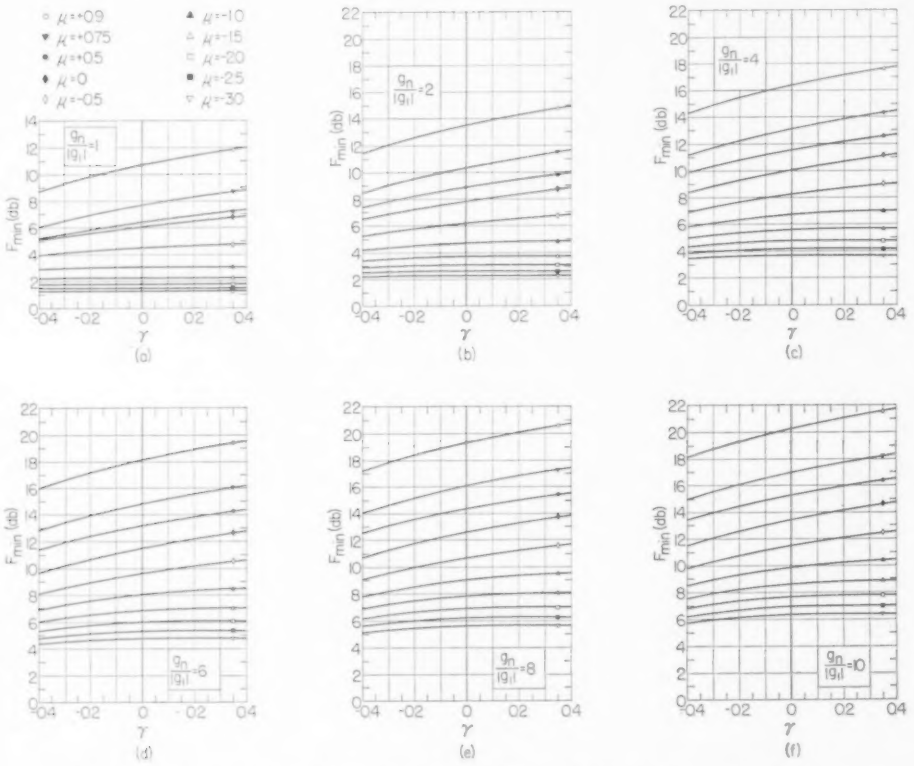


FIG. 21. Minimum total noise figure as a function of the correlation and modulation coefficients and for various values of the equivalent noise conductance. Load temperature = 300°K.

4. NUMERICAL ANALYSIS OF AN ACTUAL DIODE CHARACTERISTIC

In this Section we shall discuss the results of a numerical Fourier analysis applied to a high frequency germanium Esaki diode. The results are presented graphically as functions of the pump-operating conditions, i.e. the bias voltage and pump amplitude. The pump voltage was assumed to be sinusoidal,

$$V_p(t) = V_0 + 2V_{p1} \cos \omega_p t \qquad (4.1)$$

For this pump voltage, the bias current $I(V)$, the sum current $I_n(V)$, and the incremental conductance $g(V)$ are all periodic functions of time and can be expanded in a cosine series. The Fourier coefficients of these series are functions of the bias voltage V_0 and pump amplitude $2V_{p1}$.

The I–V characteristic was obtained with an X–Y recorder. The voltage range was divided into sections, so that the resulting expanded voltage scales would permit reading of current values at 5-mV intervals. This expansion was necessary since the incremental conductance $g(V)$ was computed from the I–V characteristic by graphical differentiation. The sum current I_n was obtained from $I(V)$ with the help of equation (3.3). Room-temperature operation was assumed.

The Fourier coefficients were calculated numerically by the well-known N-ordinate scheme.[21] In our computations, the fundamental period was divided into 15°-intervals, that is, $N = 24$, which was more than adequate for an accurate determination of the first two or three Fourier coefficients.

The germanium diode analyzed was an early version of the D4115A (Sylvania).[22] This is a high-frequency unit encapsulated in a low-inductance "pill-box" package suitable for microwave applications. The manufacturer specified a maximum (measured) oscillation frequency of 4 kMc/s. The additional information listed below was obtained at our laboratory, except where noted:

Peak voltage = 62 mV
Valley voltage = 360 mV
Peak current = 1·37 mA
Peak-to-valley ratio = 7·6
Junction capacitance $C = 3\ \mu\mu$F (Sylvania)
Series resistance $R_b = 1{\cdot}3\ \Omega$ (Sylvania)
Series inductance $L \simeq 0{\cdot}3$ mμH (estimated)
Minimum incremental conductance $g_{min} = -11{\cdot}2$ mmhos

The results of our measurements and numerical calculations are summarized in Figs. 22 and 23.

4.1 *Mixer parameters*

In Fig. 22(a) we show the portion of the I–V characteristic which was used in the analysis. This extends from the reverse-bias region to the beginning of the valley region. The calculated sum current curve, also included in this figure, shows a slight downward trend at about $V = 40$mV and decreases nearly linearly with bias until it coalesces with the $I(V)$ characteristic at approximately 100 mV. [In contrast, the curve obtained for a low-frequency gallium-arsenide diode (not shown) displays a very pronounced dip below the peak voltage, more like that sketched in Fig. 13.] The incremental conductance curve attains its minimum value $g_{min} = -11{\cdot}2$ mmhos at 115 mV.

The various mixer parameters were computed for three pump amplitudes, $2V_{p1} = 20$ mV, 50 mV, 80 mV, and for bias voltages V_0 extending from 50 mV, which is somewhat to the left of the peak voltage, to 120 mV, which is slightly beyond the g_{min} point. For the largest pump amplitude the diode is actually driven into its highly conducting reverse-bias region when biased at the lower end of the range. On the other hand, at the upper end it is pumped far into the valley region. The g_0-curves displayed in the next graph, Fig. 22(b), all exhibit the general shape of the $g(V)$ curve even though they are obtained as averages of $g(V)$ over a pump cycle. Notice that for any specified bias, g_0 decreases with decreasing pump amplitude and approaches, though never becomes less than, the incremental conductance in the limit of vanishing pump drive. Obviously, g_0 can never be less than g_{min}. The bias voltages at which $g_0 = 0$ are all to the right of the point where the incremental conductance vanishes, and move farther away with increasing pump drive.

The modulation conductance g_1 plotted in Fig. 22(c) is negative for nearly all pump conditions.

It vanishes for the lowest pump amplitude when the diode is biased near the g_{min} point. Both of these observations are in accord with our theoretical predictions. (See Subsection 1.2.) to this point, are still considerably to the left of it. The comparable size of $|g_1|$ and g_2 suggests that subharmonic pumping might prove feasible when this diode is biased near the g_{min} point.

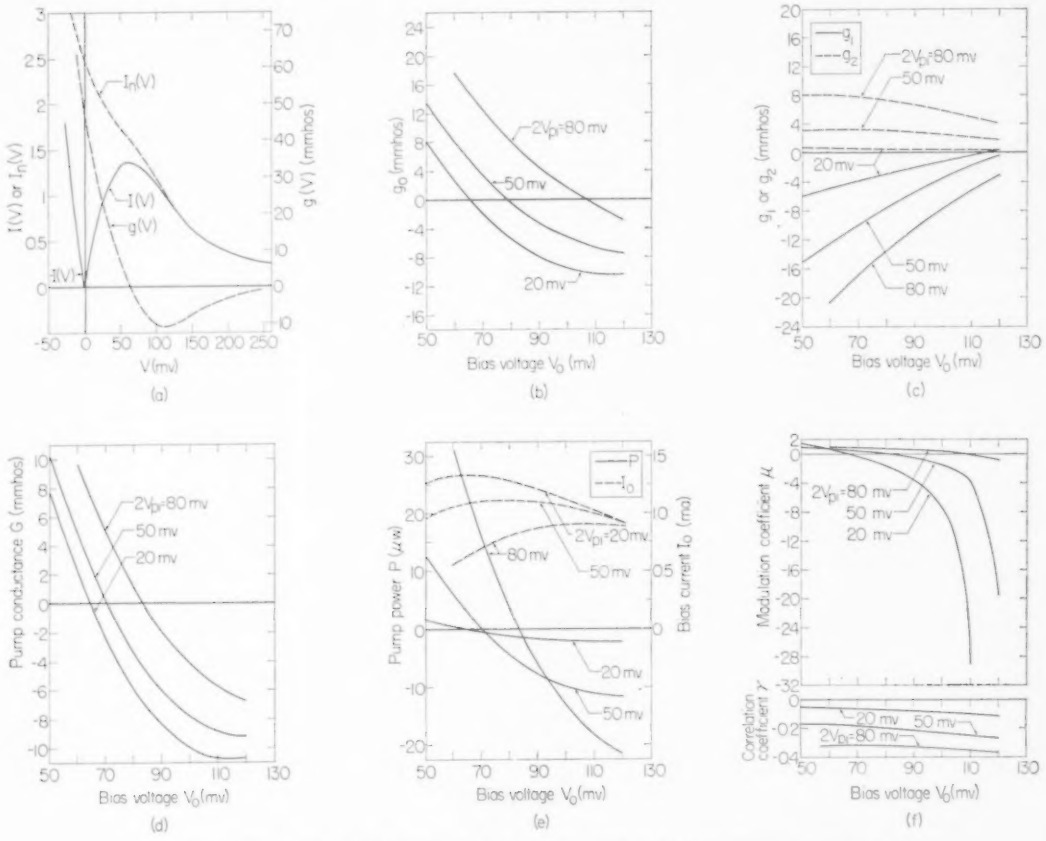


FIG. 22. Numerically calculated mixer parameters of a high-frequency germanium Esaki diode for the pump voltage $V_p = V_0 + 2V_{p1} \cos \omega_p t$.

That g_1 should decrease in magnitude with decreasing pump amplitude and be negative in sign is to be expected since it is, roughly speaking, proportional to the product of the pump amplitude and the derivative of $g(V)$.

The second harmonic coefficient g_2, also plotted in Fig. 22(c), is a positive, increasing function of pump drive and is comparable in magnitude to g_1 near the g_{min} point. For a vanishingly small pump drive, g_2 should exhibit a maximum at this point. The maxima of the g_2-curves, though tending However, for the same reason, suppression of the nearest undesired sideband frequency might become difficult. More will be said of this in the next section.

The large-signal conductance presented to the pump source is plotted in Fig. 22(d). This conductance, which we denote by G, is defined as

$$G = \frac{I_{p1}}{V_{p1}} \tag{4.2}$$

where I_{p1} is the Fourier amplitude of the

fundamental component of the pump tunnel current,

$$I_p(t) = I_0 + 2I_{p1}\cos\omega_p t + 2I_{p2}\cos 2\omega_p t \ldots \quad (4.3)$$

The pump conductance is positive for all bias voltages to the left of the peak voltage, though it becomes negative for bias voltages sufficiently to the right of this peak.

When G is negative, the loading in the pump circuit must exceed $|G|$, otherwise oscillations will occur in this circuit. On the other hand, if oscillations are desired, such as in self-excited operation, the maximum external conductance "seen" by the diode at the pump frequency must equal $|G|$. Thus curves of $G(V_0, V_{p1})$ may be used in the design of a self-excited mixer. For example, suppose that a self-excited mixer using the diode under discussion is to be designed, and suppose that a pump amplitude of 50 mV is desired when the diode is biased at 110 mV. The appropriate curve indicates that the total conductive load seen by the diode at the pump frequency should be approximately 9 mmhos. If the conductance is less than this, the diode will oscillate with a larger amplitude until stable oscillations ensue. The converse is true if the loading is too large.

The pump conductance curves resemble the small-signal conductance curves both in shape and in their relative dependence on pump amplitude. This is not coincidental, since the pump conductance can be expressed solely in terms of the (small-signal) conductances g_0 and g_2 for sinusoidal pumping conditions, as we show in the Appendix. The relation is

$$G = g_0 - g_2 \quad (4.4)$$

Since g_2 is positive, G is always more negative than g_0, though it cannot be less than g_{min}.

Curves of the diode pump power $P = 2GV_{p1}^2$ are illustrated in the next graph, Fig. 22(e). The positive values represent the power absorbed from the pump source when the diode is externally excited, the negative values correspond to the pump power fed back to the pump load when the diode is self-excited. In either case, the power levels are in the low microwatt range.

The average pump current or bias current I_0 is also plotted in Fig. 22(e). It is evident that under pumped conditions, the bias current generally decreases from its quiescent (unpumped) value $I(V)$, the decrease being greater the larger the pump amplitude.

The curves of the modulation coefficient μ, plotted in Fig. 22(f), show the same general shape for all pump amplitudes. The bias voltage where μ increases to unity corresponds to the passive limit, below which the mixer cannot exhibit gain. The correlation coefficient, also plotted in this figure, is negative over the entire range of pumping conditions and shows only a mild dependence on bias voltage. The values are all below 0·5 and increase in magnitude, approximately, as linear functions of the pump drive. This is to be expected since I_n is nearly a linear function of bias. Because γ is considerably less than unity, one may conclude that $I_n(t)$ is more nearly sinusoidal than pulsive, even though the diode is driven into the highly conducting reverse-bias region in one case. Although noise compensation occurs for practically the entire range of operating conditions considered (since g_1 is negative nearly everywhere), the degree of compensation is only moderate because of the smallness of $|\gamma|$.

The over-all mixer characteristics computed from the mixer parameters are illustrated in the next series of graphs, Fig. 23.

4.2 *Mixer characteristics*

The maximum gain-bandwidth product, Fig. 23(a), was calculated for the high-gain condition and under the assumption that only the diode capacitance was present in the signal circuits. The maximum gain-bandwidth product of the corresponding amplifier is also included for comparison.

As is clearly evident, the maximum gain-bandwidth product increases rapidly as the bias voltage enters the negative conductance region, that is, the region beyond the peak voltage. In this region, the product actually *increases* as the pump drive decreases, and, theoretically, approaches the gain-bandwidth product of the Esaki amplifier[10]

$$\frac{|g(V_0)|}{2\pi C}$$

in the limit of vanishing pump drive. That this

should be the case is evident, since in this limit $g_1 \to 0$, and $g_0 \to g(V_0)$, the incremental conductance. Of course, right *at* the limit, $g_1 = 0$ and mixing ceases; hence the gain-bandwidth is zero, by definition. Actually, one cannot let the pump drive become too small, since the small-signal assumption would be violated. At any rate, the curves imply that the mixer gain-bandwidth product at small pump drives cannot exceed the upper bound of the maximum amplifier gain-bandwidth product,

$$\frac{|g_{\min}|}{2\pi C}$$

This equals 595 Mc/s.

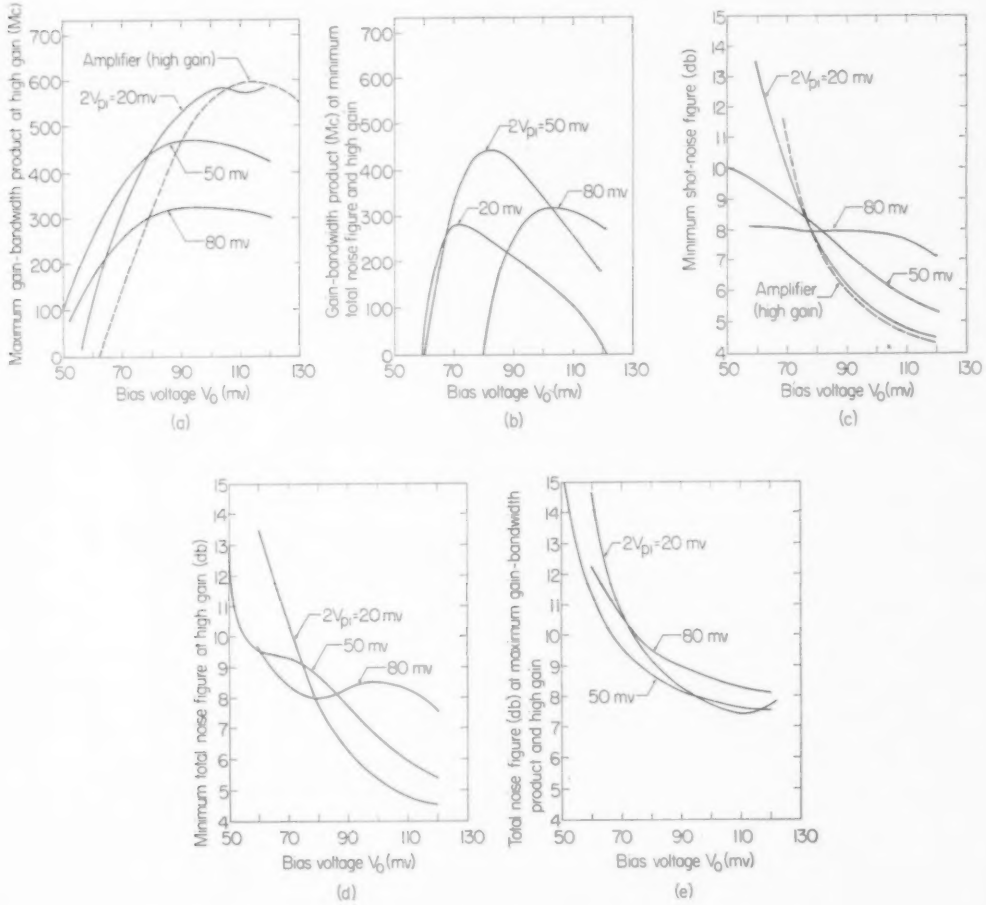


FIG. 23. Numerically calculated mixer characteristics of a high-frequency germanium Esaki diode for the pump voltage $V_p = V_0 + 2V_{p1} \cos \omega_p t$.

It is not known whether the slight dip in the curve corresponding to the lowest drive is real or simply a result of cumulative errors in the numerical analysis. Indications are that it is real, since a similar dip was obtained for a gallium-arsenide

diode. Notice that since g_1 vanishes for one point on this curve, namely, that corresponding to $V_0 \simeq 115$ mV, the curve should actually go to zero at this bias. However, this cannot be shown graphically. In practice, with losses present one should expect an easily discernible valley near this point.

In the calculation of the gain-bandwidth product it was assumed that $C_{11} = C_{22} = C$, the diode capacitance. The optimum source and load conductances are equal for this case, whereas they are not so, in general, when the total noise figure is optimized (except when $g_0 = 0$). Therefore, it is of practical interest to know how the gain-bandwidth product is degraded from its optimum value when the total noise figure is minimized instead. The curves in Fig. 23(b) show that the product still attains a maximum in the negative conductance region. The degradation is most pronounced at the lowest pump drive, so much so that the peak gain-bandwidth product is no longer obtained for this drive.

The next graph, Fig. 23(c), shows how the minimum shot-noise figure varies with pumping conditions. For a given pump excitation, this quantity exhibits a generally decreasing behavior as the diode is biased further into the negative conductance region. This is attributable to the reduction of the equivalent noise conductance g_n. Detailed examination of the calculations shows that g_n decreases almost linearly with bias and is practically independent of pump amplitude. The increase of the noise figure with pump drive is caused by the rapid increase of μ. [See equation (3.20c) and Fig. 18.]

Since the correlation coefficient is relatively independent of bias, noise compensation apparently does not play a major role in determining the shape of the noise figure curves, at least where the dependence on bias is concerned. It is evident, though, that compensation becomes more and more important as the pump drive increases. It is interesting to note that for the lowest noise figure calculated, namely 4·5 dB, at $V_0 = 120$ mV and $2V_{p1} = 20$ mV, correlation is actually detrimental since γ and g_1 differ in sign. However, since this effect is small in this case, it does not affect the noise figure measurably.

As another basis for comparison between the Esaki mixer and the Esaki amplifier we have included a curve of the latter's noise figure obtained at high gain, which is given by the expression[10]

$$F_{a_{\min}} = 1 + \frac{qI_n}{2kT_s|g|}$$

Here, as before, I_n is the sum current and g the negative incremental conductance. Observe that, as in the case of the gain-bandwidth product, the minimum amplifier noise figure is the asymptotic value of the mixer figure in the limit of vanishing pump drive. That this is reasonable (at least for the negative conductance region) is evident from equation (3.20c), since as this limit is approached, both $|g_1|$ and γ_{0s} vanish; consequently,

$$F_{s_{\min}} \to 1 + \frac{qI_{n0}}{2kT_s|g_0|}$$

Although I_{n0} and $|g_0|$ are averages over an oscillator cycle, they obviously converge to the corresponding amplifier quantities in this limit.

Note that we have not *proved* that the mixer noise figure in the negative conductance region must always exceed the amplifier noise figure, although our graphical computations strongly suggest this. Actually, we have only demonstrated that the amplifier noise figure is approached in the limit of vanishing pump drive. To prove the first statement one would need to compute the detailed analytic behavior of the various mixer parameters in the limit of vanishing pump drive, a difficult task since no reliable analytic expression exists for the diode characteristic. Furthermore, one could not justifiably restrict oneself to sinusoidal excitations in such an analysis. The computed results, nevertheless, seem to indicate that the mixer noise-figure curves, if extended further into the negative conductance region would not go below the value 4·13 dB. This is the lower bound of the minimum amplifier noise figure and is realized at $V_0 = 140$ mV.

Fig. 23(d) illustrates what happens to the noise figure when the load contribution is included. The lowest noise figures again occur in the negative conductance region and at the lowest pump drive. At the lower bias voltages there is less ordering of the curves with respect to pump drive than before. Notice that the minimum in the curve

corresponding to $2V_{p1} = 80\,\text{mV}$ occurs at $V_0 \simeq 80\,\text{mV}$, the bias value for which the pump just fails to drive the diode into the reverse direction. Similar results have been obtained with another diode. The noise figures corresponding to the two largest pump drives are probably optimistic at the extreme end of the negative conductance region since the pump excursions extend well into the valley region where the neglected $1/f$ noise might be important. This statement also applies to the corresponding shot-noise curves in the preceding graph and to the respective curves in the following graph.

The next graph, Fig. 23(e), shows how the minimum noise figure discussed above is degraded when the source and load conductances are chosen to optimize the gain-bandwidth product rather than the total noise figure. Though the lowest noise figures are still exhibited in the negative conductance region, they are considerably higher than the optimum values—except where $g_0 = 0$, at which point they coincide. The low pump-drive curves are affected most at the high end of the bias range, but the converse is true at the low end. Observe that at the upper end the noise figure does not necessarily decrease with a decrease in pump amplitude.

As we have intimated throughout the discussion of this example, results similar in all respects to those presented here were also obtained with a low-frequency gallium-arsenide diode. Because of this we feel that our results should be representative of all Esaki diodes having I–V characteristics similar to that of the diode considered here. We remind the reader, however, that the effect of the series diode impedance has been neglected, and that all sidebands have been assumed "short-circuited". When these assumptions are not fulfilled, the results can be seriously affected.

4.3 *Conclusions*

The most important conclusion to be drawn from the preceding numerical analysis is that at high gain the lowest noise figure and the highest gain-bandwidth product are realized when the diode is operated in the negative conductance region. We must emphasize, however, that the terminating conductances necessary for realizing these optimum situations differ in general. Consequently, in practice one must optimize one of these mixer properties at the expense of the other. Which course is adopted depends on the system requirements.

5. MISCELLANEOUS TOPICS

In Section 1, for reasons of analytic simplicity we neglected the diode series impedance and assumed that the undesired sideband signals were terminated in short circuits. We now consider the changes in the analysis that are necessary when these assumptions are not valid. We shall also touch upon other topics such as the sensitivity of the mixer gain to small variations in the mixer parameters and terminating conductances and the effect of incomplete isolation of the signal and pump circuits.

5.1 *Series diode impedance*

It is a simple matter to correct the small-signal circuit to include the effect of the series impedance. Since each sideband current must pass through this impedance, one need merely add this impedance in series with each signal and sideband port, regardless of the number of sideband components considered. Thus, if all possible components are included, the small-signal representation of the mixer, including series impedance effects, takes the form shown in Fig. 24. If R_b is frequency sensitive, the appropriate frequency-dependent value must be used at each port.

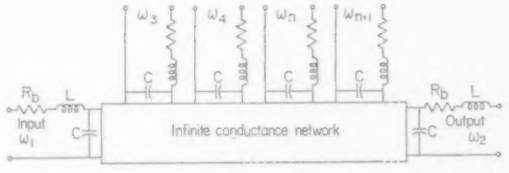


FIG. 24. Small-signal equivalent circuit of a mixer which includes the effects of the diode series impedance and all sidebands.

In order to proceed with the mixer analysis, one must specify how the various sideband ports are terminated. At this stage we arrive at an impasse. Because of the series impedance one cannot assume short-circuited sidebands. On the other hand, because of the junction capacitance one cannot assume open-circuited sidebands,

either. However, unless one or the other of these assumptions is made in order to dispose of the infinite number of sidebands, the solution to the mixer problem is unattainable. At best, we are forced to strike a compromise which will at least indicate how the solutions obtained for the idealized model change when the series impedance is included. The compromise we shall adopt, and the one that is usually made in conventional crystal mixer theory,[23,24] assumes that all but a finite number of sidebands are shorted, as in the ideal case. Actually, this assumption may have more justification in our present analysis, since the junction capacitance of an Esaki diode is generally higher than that of a comparable conventional mixer diode; therefore, the higher sidebands are more likely to see a termination approaching a short circuit. Nevertheless, there are inconsistencies in this compromise and any conclusions based on it will be semi-quantitative in nature, particularly if the series impedance level is comparable to the impedance level of the idealized diode.

In practice it is often permissible to neglect the inductance at the lower of the two signal frequencies. For example, at 30 Mc/s, a commonly used output (i.f.) frequency, the inductive reactance of a typical high-frequency Esaki diode package ($L \simeq 0{\cdot}5$ mμH) is approximately $0{\cdot}1\ \Omega$. The series resistance is much higher. On the other hand, the reactance at the higher (r.f.) signal frequency, say 1 kMc/s, is $3\ \Omega$, a value comparable to the series resistance which may range typically, from $0{\cdot}5$ to $2\ \Omega$. If we include the series impedance in the signal circuits, but assume that all undesired sideband voltages are short-circuited, we may compute the various mixer characteristics such as gain, gain-bandwidth, etc., as before. Although the calculations are straightforward, the resulting expressions are quite unwieldy and not very illuminating.[23] Unless one carries out extensive numerical computations, it is difficult to draw any significant quantitative conclusions from them. These facts, coupled with the approximate nature of the results, make it impractical to perform these calculations at the present time unless experiments show that series impedance effects are very important (and they may very well be with present-day Esaki diodes).

In this paper we shall concern ourselves only with some qualitative predictions. Since the diode inductance appears in series with the external tuning inductance rather than in shunt, the power-gain versus frequency characteristic may not exhibit a single-peaked response as before. The principal effect of the series resistance, aside from the loading of the tuned circuits, is its degradation of the noise performance because of the thermal noise contributions arising from it. This thermal noise is contributed not only by the noise in the frequency bands centered about the desired signal frequencies, but also by the noise bands centered about the frequencies of the assumed shorted sidebands. This is also true for the shot noise.

A practical consideration is the effect of the series resistance on the upper frequency limit of active mixer operation. It is well known that the series resistance imposes an upper frequency limit on the Esaki diode amplifier. This upper limit, called the resistive cut-off frequency, is the frequency above which the diode no longer exhibits an incremental negative conductance.[10] Unfortunately, one cannot extrapolate this result in a concise manner to the mixer case. This is because the mixer conductances cannot be related in a simple manner to the minimum incremental conductance used in the computation of the cut-off frequency. A further complication is that at least two signal frequencies (one usually very low) are involved in a mixer, rather than one as in the amplifier. About all that can be said is that the higher the resistive cut-off frequency, the higher the upper frequency limit of the corresponding mixer is likely to be, though even this cannot be stated with absolute certainty nor proven simply. It is quite probable, however, that the mixer will have low gain and a very high noise figure when one of the signal frequencies is near the cut-off frequency.

If the series impedance is included in the signal circuits, consistency demands that it also be included in the pump circuit. In this case, because of the harmonic voltage drops across the series impedance, it is no longer permissible to assume that the diode conductance is pumped sinusoidally even though the pump voltage impressed across the packaged diode is sinusoidal.

5.2 *Image terminations*

Experience with conventional crystal mixers has demonstrated that the assumption that all

undesired sideband components can be terminated in short circuits or near short circuits is often satisfactory (if the series impedance is not too high). The possible exception is the sideband nearest the higher of the two signal frequencies. This sideband, which is usually called the "image", is separated from the higher signal frequency by twice the lower frequency. For example, in a non-inverting down-converter, $\omega_1 > \omega_p > \omega_2$, the image frequency ω_3 is equal to $\omega_p - \omega_2 = \omega_1 - 2\omega_2$. Mixer operation with an image is often referred to as double-sideband operation to distinguish it from the single-sideband operation considered earlier. The image will be particularly important when its frequency is so near the signal frequency that circuit termination of these two components cannot proceed independently. For example, this is the case if the operating input bandwidth† (of a down-converter) is larger than four times the output frequency.

To study the effects of incomplete image suppression we postulate at the outset that the image port is terminated in an arbitrary admittance which we later shall specialize. Though for simplicity we assume that the series impedance is negligible, the analysis may be corrected in the manner described in the previous Subsection if this is not true. For concreteness we shall consider the non-inverting down-converter. If V_3 and I_3 represent the complex voltage and current amplitude of the image and if the time origin is chosen to make g_1 real, we obtain the set of equilibrium equations

$$\left.\begin{aligned} I_1 &= (g_0+j\omega_1 C)V_1+g_1V_2+g_2V_3^* \\ I_2 &= g_1V_1+(g_0+j\omega_2 C)V_2+g_1V_3^* \\ I_3^* &= g_2^*V_1+g_1V_2+(g_0-j\omega_3 C)V_3^* \end{aligned}\right\} \qquad (5.1)$$

The complex conjugates of I_3 and V_3 and the minus sign preceding ω_3 appear because the image is a lower sideband (i.e. $\omega_3 = \omega_p - \omega_2$).

The equilibrium equations can be represented by the equivalent circuit shown in Fig. 25(a). If the conductance is an even function of time, g_1 and g_2 are real and the simpler circuit of Fig. 25(b) results. These circuits reduce to the single-sideband circuit when the image terminals are shorted.

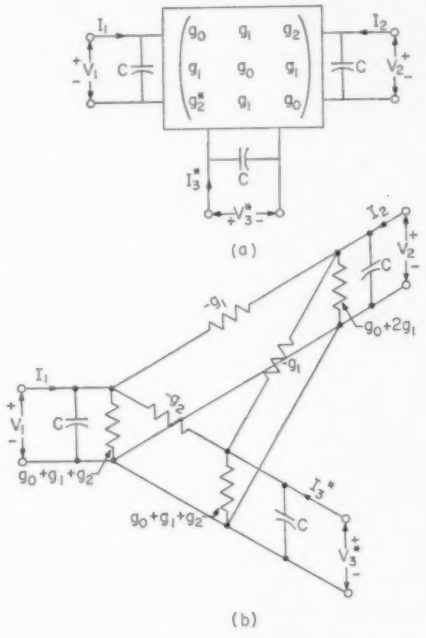


FIG. 25. Small-signal equivalent circuit of a (a) non-reciprocal and (b) reciprocal, non-inverting down-converter which includes the image sideband.

These equations show that the image interacts with the input signal via the second harmonic of the time-varying conductance, i.e. through g_2, and with the output signal via the first harmonic. When g_2 is comparable in magnitude to g_1, such as when the diode is biased very near the minimum conductance point (see Subsection 1.2) the image component might prove particularly troublesome if not adequately suppressed. Difficulties of this nature have been reported in the literature. If g_2 dominates g_1, it might be wiser to achieve mixing by using the second harmonic of $g(t)$ rather than the fundamental. This can be done by simply halving the initial pump frequency. Though the image-signal frequency separation remains unchanged in this operation, the corresponding interaction is probably small because it is proportional to the fourth harmonic coefficient

† The operating bandwidth will be less than the "cold" circuit bandwidth because of the negative input conductance of the mixer.

g_4. However, interaction between the image and output signal may not be negligible since it is proportional to g_2.

If the image port is terminated in an admittance Y_3, the three equilibrium equations can be reduced to two which pertain to the desired input and output signals alone. The elements of the resultant 2×2 conversion matrix are given by

$$\left.\begin{aligned} y_{11} &= g_0+j\omega_1 C-\frac{|g_2|^2}{g_0-j\omega_3 C+Y_3^*} \\ y_{22} &= g_0+j\omega_2 C-\frac{g_1^2}{g_0-j\omega_3 C+Y_3^*} \\ y_{12} &= g_1\left(1-\frac{g_2}{g_0-j\omega_3 C+Y_3^*}\right) \\ y_{21} &= g_1\left(1-\frac{g_2^*}{g_0-j\omega_3 C+Y_3^*}\right) \end{aligned}\right\} \quad (5.2)$$

Comparing these expressions with those for the single-sideband case, equation (1.14), one can readily see that the real parts of the admittances are no longer frequency independent. If the image-input signal interaction is small in comparison to the image-output interaction, that is, if $|g_1|\gg|g_2|\simeq 0$, the principal effect of the finite image termination is on the output admittance element y_{22}. On the other hand, if the opposite is true, y_{11} is affected more than y_{22}, as are the transfer admittances y_{12} and y_{21}. When g_2 is real, such as for sinusoidal pumping, the mixer is reciprocal, that is $y_{12}=y_{21}$. We shall assume this to be the case from here on.

Because of the many possible image terminations that could be considered, space does not permit us to treat the image case in complete generality or in detail, especially in regard to the frequency response of the transducer gain. Therefore we shall briefly sketch how some of the more important results obtained for the single-sideband case are affected when the image is terminated in certain special ways that may occur in practice. In particular, we shall consider only the case when the image is at resonance at the image center frequency $\Omega_3=\omega_p-\Omega_2$. At resonance, the image circuit contributes a conductive term to the two-port admittances. In particular, the real parts of these admittances become

$$g_{11}=g_0-\frac{g_2^2}{g_0+g_i+g'''} \quad (5.3a)$$

$$g_{22}=g_0-\frac{g_1^2}{g_0+g_i+g'''} \quad (5.3b)$$

$$g_{12}=g_{21}=g_1\left(1-\frac{g_2}{g_0+g_i+g'''}\right) \quad (5.3c)$$

Here g_i represents any externally applied dissipative load at the image terminals, and g''' any parasitic losses.

Two special image loads will be considered, namely,

(a) $\qquad g_i=0$

which we shall designate as the open-circuited case, and

(b) $\qquad g_i=g_s$

the so-called "broadband" situation to which we alluded earlier. This latter case, a practical one, is approximately realized whenever the image and nearby input signal frequency are so close to each other that the input tuned circuit cannot separate the two frequencies. In other words, the image and input terminals of the converter element "see" approximately the same circuit admittance (at least over a band about resonance). Consideration of the above two image loads coupled with the previous results for the shorted-image case should provide us with a fair measure of the effect of intermediate image loads.

Open-circuited image case. The general expression for midband gain for this case is

$$K_0=\frac{4g_s g_l g_{12}^2}{[(g_s+g'+g_{11})(g_l+g''+g_{22})-g_{12}^2]^2} \quad (5.4)$$

where g_{11}, g_{22} and g_{12} are obtained from equation (5.3) by setting $g_i=0$.

The stability conditions require that

$$\left.\begin{aligned} &g_s+g'+g_{11}>0 \\ &g_l+g''+g_{22}>0 \\ &(g_s+g'+g_{11})(g_l+g''+g_{22})>g_{12}^2 \end{aligned}\right\} \quad (5.5)$$

Comparing these conditions with those for the single-sideband case, one can easily show that a

mixer, stable under short-circuited image conditions, may oscillate under open-circuited image conditions if the loading at the signal ports is insufficient. Because of the frequency dependence of g_{11}, g_{22}, y_{12} and y_{21} off resonance, it cannot be stated without further calculations that the above conditions are sufficient to guarantee stability off resonance.

In order to prove that gain can be obtained at resonance, we must apply the general activity conditions equation (1.15) to the mixer admittances. If image losses are negligible, $g''' \simeq 0$, it can be shown that the mixer will be active *unless* the second harmonic coefficient g_2 satisfies the inequalities

$$g_0\left[2\left(\frac{g_1}{g_0}\right)^2 - 1\right] < g_2 \leqslant g_0$$

Study of this restriction shows that a mixer can be active under open-circuited image conditions even though it is not active under short-circuited image conditions, i.e. even if the inequality $g_0 < |g_1|$ is *not* fulfilled. However, it is still necessary that a region of negative slope exist in the operating range of the pump. However, if the latter inequality *is* fulfilled, the restrictive relation above predicts that the mixer will be active for all values of g_2 provided that g_0 does not fall in the range

$$-|g_1| < g_0 < 0$$

Turning attention to the noise properties, one can expect that the noise performance will be worse than that of the single-sideband mixer for comparable terminating conditions. The reason for this is that the image circuit contributes at least two additional sources of noise, a thermal source representing the image load and the image losses and a shot source corresponding to the image band shot noise. Both sources can be represented by current generators.

In calculating the various noise power terms we note that, in addition to the existence of correlation between the frequency-translated counterpart of the input shot-noise generator n_1 (i.e. n_{12}) and the output generator n_2, there is also correlation between these two generators and the frequency-translated image generator n_3, which we denote by n_{32}. Thus, there are six shot-noise contributions to the output noise power: three self-power terms

$$\overline{\langle n_{12}^2 \rangle} = \overline{\langle n_{32}^2 \rangle} = \overline{\langle n_2^2 \rangle} = 2qI_{n0}\,\delta f \qquad (5.6a)$$

and three correlation terms

$$\overline{\langle n_{12} n_2 \rangle} = \overline{\langle n_{32} n_2 \rangle} = 2qI_{n1}\,\delta f \qquad (5.6b)$$

$$\overline{\langle n_{12} n_{32} \rangle} = 2qI_{n2}\,\delta f \qquad (5.6c)$$

In other words, the image circuit is responsible for three shot terms, i.e. one self-power term and two correlation terms. The correlation term involving n_3 and n_2 is proportional to the first harmonic amplitude of the sum current. On the other hand, the other correlation term, representing the interaction between the input and image generators, is proportional to the second harmonic amplitude. The reason for this is that in a down-converter, interaction between the input signal and the image occurs via the second harmonic of the time-varying conductance as we explained earlier.

After some rather tedious algebraic gymnastics we arrive at the following expression for the total noise figure:

$$F = 1 + \frac{g_n}{g_s}[(\gamma A - \gamma' - B)^2 - (\gamma A - \gamma')^2 + (A - \gamma)^2 + 1 - \gamma^2] + \frac{g'T'}{g_s T_s} + B^2\left(\frac{g'''T'''}{g_s T_s}\right) + A^2\left(\frac{g_l T_l + g''T''}{g_s T_s}\right) \qquad (5.7)$$

where

$$A = \frac{(g_0 + g''')(g_s + g_0 + g') - g_2^2}{g_1(g_0 + g''' - g_2)}$$

$$B = \left(\frac{g_s + g' + g_0 - g_2}{g''' + g_0 - g_2}\right)$$

and

$$\gamma' = \frac{I_{n2}}{I_{n0}}$$

This last quantity represents the correlation between the image and input shot noise generators.*

* It is worth mentioning here that the two correlation coefficients γ and γ' are not entirely independent of each other. It can be shown that because of the non-negativeness of the tunnel sum current, they must satisfy the inequality $2\gamma^2 - 1 \leqslant \gamma' \leqslant 1$.

In the above expression the terms in brackets denote the shot-noise contributions. The three terms following this represent the various thermal contributions. When the image loss conductance becomes arbitrarily large, F approaches the expression for the shorted-image case equation (3.14).

Broadband case. As mentioned earlier this case is distinguished by the fact that, to a first approximation, the circuit admittance seen by the image port is the same as that seen by the input port. Though this is true, it is not permissible to connect these ports in parallel for computational purposes since they still pertain to two distinct frequencies. Instead, one must specialize equation (5.2) to this case by setting Y_3 equal to the source termination Y_1. At resonance the equivalent two-port admittances become

$$\left.\begin{aligned} g_{11} &= g_0 - \frac{g_2^2}{g_0+g_s+g'} \\ g_{22} &= g_0 - \frac{g_1^2}{g_0+g_s+g'} \\ g_{12} &= g_{21} = g_1\left(1-\frac{g_2}{g_0+g_s+g'}\right) \end{aligned}\right\} \tag{5.8}$$

Since the mixer conductances depend on the source conductance, it is not convenient to derive the activity conditions in the manner employed in all previous cases. Rather, we shall deduce these conditions from the power gain expression. The transducer gain at resonance is given by

$$K_0 = \frac{4g_s g_l g_1^2}{[(g_s+g'+g_0+g_2)(g_l+g''+g_0)-2g_1^2]^2} \tag{5.9}$$

The mixer will be active provided that at least one of the inequalities

$$g_0 < 0 \tag{5.10a}$$

$$g_0+g_2 < 0 \tag{5.10b}$$

$$g_0(g_0+g_2) < 2g_1^2 \tag{5.10c}$$

is satisfied.* These conditions are equivalent to the statement that the mixer will be active for all values of g_0, g_1 and g_2 *except* those which satisfy the *pair* of conditions

(a) $$g_0 > 0$$

(b) $$g_2 > g_0\left[2\left(\frac{g_1}{g_0}\right)^2 - 1\right]$$

These inequalities are very similar, but not identical to the conditions obtained for the open-image case; however, they still rule out the possibility of obtaining gain with a positive nonlinear conductance.

The two-port stability criteria are not directly applicable to the broadband case because the equivalent two-port matrix elements are not independent of the external (source) loading, an implied requirement in the derivation of the two-port stability criteria. To treat the stability problem for this case we need to extend the stability criteria to the three-port. However, even then some difficulties arise because the loading presented to two of the ports are not independent. Because of these difficulties we shall simply make some studied guesses of the stability requirements by inspecting the denominator of the gain expression. By analogy with the results for the shorted-image case, it is *probably* true that for stability it will be necessary that the conditions

$$g_s+g'+g_0+g_2 > 0 \tag{5.11a}$$

$$g_l+g''+g_0 > 0 \tag{5.11b}$$

$$(g_s+g'+g_0+g_2')(g_l+g''+g_0) > 2g_1^2 \tag{5.11c}$$

be fulfilled. It is not known what other conditions must be satisfied.

The noise figure may be obtained from the open-image formula by merely letting

$$g''' \rightarrow g'+g_s$$

$$g'''T''' \rightarrow g'T'+g_sT_s$$

These transformations lead to the following changes

$$A \rightarrow \frac{g_s+g'+g_0+g_2}{g_1}$$

$$B \rightarrow 1$$

* This can be proven by showing that if any of these conditions hold, it is always possible to choose a pair of passive terminations $g_s+g'>0$, $g_l+g''>0$, which will make the denominator arbitrarily small.

Hence,

$$F = 2\left(1+\frac{g'T'}{g_sT_s}\right) + \frac{g_n}{g_s}\left[(A-\gamma)^2+(1-\gamma^2)+1-2\gamma A+2\gamma'\right] + A^2\left(\frac{g_lT_l+g''T''}{g_sT_s}\right) \tag{5.12}$$

The first terms in parentheses represent the thermal noise contributed by the source and by the input loss conductance, and the last terms represent the corresponding noise of the load and the output loss conductance. The terms in brackets denote the shot noise. The factor 2 preceding the input thermal contributions accounts for the fact that the source and the input loss conductance contribute equal amounts of noise through the input and image terminals. In other words the noise figure given above is the one that would be measured using a narrow-band noise source, and therefore cannot be less than 2 (3 dB). It is not to be confused with the double-sideband noise figure which is measured using a broadband source. This latter quantity is obtained by dividing equation (5.12) by 2.

Comparing the shot terms with those for the shorted-image case, it is seen that the first two terms correspond to the shot contributions obtained for the latter case. The remaining three terms are attributable to the image circuit alone. For example, the unity term in brackets corresponds to the self-power shot noise of the image circuit, the term $-2\gamma A$ represents the correlation between the image and output shot noise [it matches an equal contribution due to the signal-output correlation which is included in the term $(A-\gamma)^2$], and, finally, the term $2\gamma'$ denotes the correlation between the input and image shot noise. It is interesting to note that the shot noise contribution can never be less than $(A-2\gamma)^2$. This is easily proven by using the inequality $\gamma' \geqslant 2\gamma^2-1$.

5.3 *Effects of the pump circuit*

Unless properly isolated, the pump circuit will load the signal circuits. If the pump circuit presents admittances y_{p1} and y_{p2} at the diode terminals at the frequencies ω_1 and ω_2, respectively, these admittances will appear directly across the respective ports of the small-signal circuit. These admittances, if purely imaginary, will be detrimental only if they are capacitive in nature, since the gain-bandwidth product will be reduced. On the other hand, they will always be detrimental if there is a real part associated with either, since this will reduce the gain and introduce thermal noise. Both of these effects can be taken into consideration simply by including the real parts g_{p1} and g_{p2} as components of the input and output loss conductances. Some loading may be unavoidable if the output frequency of the converter is small compared to the input frequency, since then the latter is nearly equal to the pump frequency. In this case, it is essential to include the effects of pump loading, unless the pump is loosely coupled to the diode. However, it may be undesirable to use loose coupling, since this can introduce harmonics into the pump voltage.

A case when the pump circuit cannot be loosely coupled is when the diode is self-excited. In order to avoid the additional thermal noise of the pump load in this situation, it may be advantageous to stabilize the pump oscillations with the signal circuit loads. However, this procedure has several disadvantages, one being that large pump voltages appear in the signal circuits. Another is that tuning of the signal circuits or the presence of large signal voltages may affect the oscillation frequency (pulling).

In addition to the thermal noise of the pump circuit, noise of a non-thermal origin may also be introduced, particularly when an external pump is used. However, one can minimize this noise by inserting a narrow-band filter between the pump and mixer. Alternatively, various noise cancellation schemes may be employed.[25]

5.4 *Gain sensitivity*

The gain of a negative conductance mixer, like that of the corresponding amplifier, is highly sensitive to small fluctuations of the device conductances and of the terminations, especially at high gain. This sensitivity can be stated in terms of the fractional change in gain per fractional change in a parameter. For example, the sensitivities of the two-port mixer to small changes in the mixer conductances g_0 and g_1 and in the

source conductance g_s are, respectively,

$$S_0 = \frac{\delta K_0/K_0}{\delta g_0/g_0} = -\left[\frac{g_0}{|g_1|}\left(\frac{2g_0+g_s+g_l}{\sqrt{(g_s g_l)}}\right)\right]\sqrt{(K_0)} \tag{5.13a}$$

$$S_1 = \frac{\delta K_0/K_0}{\delta|g_1|/|g_1|} = 2\left[1+\frac{2|g_1|}{\sqrt{(g_s g_l)}}\right]\sqrt{(K_0)} \tag{5.13b}$$

$$S_s = \frac{\delta K_0/K_0}{\delta g_s/g_s} = \left[1-\sqrt{\left(\frac{g_s}{g_l}\right)}\left(\frac{g_l+g_0}{|g_1|}\right)\sqrt{(K_0)}\right] \tag{5.13c}$$

The sensitivity to load fluctuations is identical to the latter except that g_s and g_l are interchanged.

In all cases the sensitivities at high gain increase as $K_0^{1/2}$, a behavior typical of negative conductance devices. For example, when

$g_s = g_l$,

$$S_s \simeq 1-\sqrt{(K_0)}$$

at high gain. Thus, if $K_0 = 100$, a 1 per cent change in the source conductance results in 9 per cent change in gain. Since the source admittance presented to a microwave mixer is rarely constant, this points out a problem that may be important in practice. Likewise, sensitivities to variations in the device conductances emphasize the need for a stable bias and pump source.

SUMMARY

We have shown that because of the nonlinear negative conductance of the Esaki diode, frequency conversion with gain is possible if this conductance is varied periodically by a large pump voltage. When only two small signal voltages are present, the signal conversion properties of the resultant linear mixer can be characterized by a simple 2×2 matrix, the elements of which are certain Fourier coefficients of the periodic series representing the time-varying conductance. These elements are independent of the two signal frequencies, hence are the same for up-conversion and down-conversion.

In the practical case, when the difference of the two signal frequencies is equal to the fundamental pump frequency, the matrix elements are equal to the first two Fourier coefficients, namely, g_0, the average value of the time-varying conductance and g_1, the amplitude of the first harmonic. The latter plays the dominant role in the conversion process. Either of these conductances may be positive or negative. It was shown that conversion with gain is possible only if $g_0<|g_1|$. When this is true, arbitrarily high gain is theoretically possible. We have also demonstrated that stable mixer operation can be obtained provided that the source and load conductances exceed a certain lower bound specified by g_0 and g_1 and the terminal conductances themselves.

When $g_0<0$, the mixer can only present a negative input and output conductance at resonance. On the other hand, when $g_0>0$ it can also present a positive conductance at either port. However, these conductances cannot be positive simultaneously unless the resonance gain is less than 6 dB.

Formulas were derived for the gain-bandwidth product and the noise figure, and for the loading conditions at the two signal ports which optimize these two quantities at high gain. It was shown that under optimal conditions the product of the bandwidth and the square root of the midband gain is independent of gain and is only a function of the mixer parameters. Specifically, it is proportional to $|g_1|-g_0$ and inversely proportional to C, the diode junction capacitance. At a low gain, the product may be higher.

We have demonstrated that the ultimate noise performance of the Esaki mixer is determined by the diode shot noise. It was also shown that under proper conditions, partial compensation of the shot noise may result because of correlation effects. The conditions necessary for this to happen were considered in some detail.

With the aid of results obtained from a numerical Fourier analysis applied to an actual Esaki diode characteristic, we have demonstrated that under sinusoidal pumping conditions the most optimum mixer operation, based on gain-bandwidth and noise considerations alone, is realized when the diode is pumped very lightly and biased in its negative conductance region. For this situation the calculated results imply that at high gain the gain-bandwidth product is always lower than, and the noise figure is higher than, the most optimum gain-bandwidth product and noise figure of the corresponding Esaki amplifier operated at high gain.

Since the terminating conditions corresponding to the optimum gain-bandwidth product and

optimum noise figure do not coincide, except in certain special cases, a compromise will be necessary in any specific application.

CONCLUSIONS

In comparing the relative merits of operating the mixer under the two possible conditions, $g_0 > 0$ or $g_0 < 0$, we are aided by the results of the numerical analysis. These have shown us that operation under the latter condition is definitely preferred whenever gain-bandwidth or noise considerations are of prime importance. However, other system requirements may rule against this operation. For example, if the mixer must present a positive input or positive output conductance, or both, only the former operating condition is permissible. However, even for this condition, the gain is at best less than 6 dB if both port conductances must be positive. Another basis for comparison centers about the manner of pump excitation. Thus, although self-excitation is possible for both conditions, the permissible range of pump amplitudes and bias voltages is more limited when $g_0 > 0$. Furthermore, this condition is less attractive from the standpoint of stability of the self-oscillations. On the other hand only this operating condition permits external excitation without requiring heavy external loading in the pump circuit to prevent spurious oscillations.

There is a much broader basis for comparison when the relative advantages and disadvantages of Esaki mixers and conventional crystal mixers are considered. For example, the possibilities of gain and of self-excitation are two distinct points in favor of Esaki mixers. The low pump power requirement for external excitation is another. For typical operating conditions conventional mixers require about 0·5–1·0 mW of pump power.[25] Esaki mixers can operate at power levels at least an order of magnitude lower than this, as our example has shown. This is particularly true when the diode is biased near the peak current point. However, d.c. power requirements are not significantly lower for Esaki mixers, being of the order of 50–100 μW.

Because of the tunneling mechanism and high-doping levels associated with them, Esaki diodes are less sensitive to radiation damage, surface contamination and temperature fluctuations than conventional diodes. On the other hand, when the insensitivity to fluctuations of the pumping conditions and signal terminations is compared, the conventional crystal mixer must be ruled superior.

Another advantage of the conventional mixer is the fact that positive conductances are presented to both ports under all operating conditions.* This is also possible with Esaki mixers, but only if the gain is less than 6 dB.

Conventional crystal mixers are also superior in their ability to maintain linear operation at considerably larger signal input levels. There are several reasons for this. For linear operation the total signal voltage across the diode must be a small fraction of the pump voltage. Because of the relatively small voltage range suitable for mixer operation in Esaki diodes, the maximum pump amplitudes permitted for optimum operation are much lower than for conventional diodes. Furthermore, for the same maximum permissible output, the signal input power that can be handled by the Esaki mixer is less than that for the conventional mixer by a factor approximately equal to the gain. This is appreciable when the gain is high.

When noise figures are considered, the comparison is far from decisive. Mixers using present-day Esaki diodes operated under the most optimum conditions are theoretically capable of exhibiting noise figures superior to the best achieved with conventional mixers. However, the practical realization of the optimal conditions is another matter. Except for the results of DICKENS and GNEITLING[7] and CHANG *et al.*[5], the reported noise figures to date are not exceptional. Whether this situation will be changed appreciably when more favorable experimental conditions are realized or when better diodes become available remains to be seen.

APPENDIX

The amplitude of the fundamental component of pump tunnel current is given by

$$I_{p1} = \frac{1}{2\pi} \int_0^{2\pi} I(V_p) \cos t \, dt \qquad \text{(A.1a)}$$

* This is true, if parametric effects arising from the nonlinear junction capacitance are negligible.

Integrating by parts,

$$I_{p1} = \frac{1}{2\pi}\left[I(V_p)\sin t\Big|_0^{2\pi} - \int_0^{2\pi}\frac{\partial I}{\partial t}\sin t\,dt\right] \quad \text{(A.1b)}$$

The partial derivative is taken with V_0 and V_{p1} held constant. The first term vanishes, obviously. Since $V_p = V_0 + 2V_{p1}\cos t$ the second term can be transformed with the help of the identity

$$\frac{\partial I(V_p)}{\partial t} = \frac{\partial V_p}{\partial t}\frac{\partial I(V)}{\partial V}\Bigg|_{V=V_p}$$
$$= -2V_{p1}g(V_p)\sin t$$

Inserting this in equation (A.1b) and using the identity $2\sin^2 x = (1-\cos 2x)$, we obtain

$$I_{p1} = \frac{V_{p1}}{2\pi}\int_0^{2\pi} g(V_p)[1-\cos 2t]dt \quad \text{(A.2)}$$

hence

$$\frac{I_{p1}}{V_{p1}} = G = g_0 - g_2 \quad \text{(A.3)}$$

Acknowledgements—It is with great pleasure that I acknowledge the help of the many individuals who co-operated in the work reported in this paper. I am particularly grateful to Miss BARBARA KILEY who performed most of the calculations. I wish to thank my colleagues, Dr. JAMES MULLEN who gave helpful suggestions on certain aspects of the noise problem, also Dr. HAROLD ROTH, Mr. PHILLIP ZELDES and Mr. ANTHONY SCHMID who were generous with experimental assistance on various occasions.

REFERENCES

1. F. E. TERMAN, *Radio Engineering* Chap. 10. McGraw Hill, New York (1947).
2. G. C. MESSINGER and C. T. MCCOY, *Proc. I.R.E.* **45**, 1269 (1957).
3. R. RAFUSE, Thesis, M.I.T. (1960).
4. L. ESAKI, *Phys. Rev.* **109**, 603 (1958).
5. K. K. N. CHANG, G. H. HEILMEIER and H. J. PRAGER, *Proc. I.R.E.* **48**, 854 (1960).
6. C. S. KIM and J. E. SPARKS, *Proc. Nat. Electron. Conf.* **16**, 791 (1960).
7. L. E. DICKENS and C. R. GNEITLING, *I.R.E. Trans.* **MTT-9**, 99 (1961).
8. L. ESAKI and Y. MIYAHARA, *Solid-State Electron.* **1**, 13 (1960).
9. E. F. BOLINDER, *I.R.E. Trans.* **CT-4**, 70 (1957).
10. R. A. PUCEL, *Solid-State Electron.* **1**, 22 (1960).
11. T. J. YAJIMA and L. ESAKI, *J. Phys. Soc. Japan* **13**, 1281 (1958).
12. R. A. PUCEL, *Proc. I.R.E.* **49**, 1080 (1961).
13. W. B. DAVENPORT and W. L. ROOT, *An Introduction to the Theory of Random Signals and Noise* Chap 7. McGraw-Hill, New York (1958).
14. A. UHLIR, *Bell Syst. Tech. J.* **37**, 951 (1958).
15. M. J. O. STRUTT, *Proc. I.R.E.* **34**, 942 (1946).
16. A. VAN DER ZIEL, *Noise* Chap. 9. Prentice-Hall, Englewood Cliffs, N.J. (1954).
17. S. O. RICE, *Bell Syst. Tech. J.* **23**, 282 (1944); **24**, 46 (1945).
18. A. BLANC-LAPIERRE and R. FORTET, *Théorie des Fonctions Aléatoires* Chap. 5. Masson, Paris (1953).
19. D. I. BREITZER, *Proc. I.R.E.* **48**, 935 (1960).
20. G. E. VALLEY and H. WALLMAN, *Vacuum Tube Amplifiers* Chap. 13. McGraw-Hill, New York (1948).
21. S. WHITTAKER and G. ROBINSON, *The Calculus of Observations* Chap. 10. Blackie, London (1944).
22. G. DERMIT, H. LOCKWOOD and H. HAUER, *Proc. I.R.E.* **49**, 519 (1961).
23. L. C. PETERSON and F. B. LLEWELLYN, *Proc. I.R.E.* **33**, 458 (1945).
24. H. TORREY and C. WHITMER, *Crystal Rectifiers* Chap. 5. McGraw-Hill, New York (1948).
25. R. V. POUND, *Microwave Mixers.* McGraw-Hill, New York (1948).

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 208–214. Printed in Great Britain

TUNNEL-DIODE POWER

G. DERMIT

General Telephone and Electronics Laboratories, Inc., Bayside 60, N.Y.

(*Received* 24 *March* 1961; *in revised form* 7 *June* 1961)

Abstract—A derivation of an expression for nearly linear tunnel-diode power is given. This expression is based on the fact that the time-constant *RC* of tunnel junctions is a sensitive function of the impurity concentration in the semiconductor material, whereas the series resistance–capacitance product is a very slowly varying function of impurity concentration. Tunnel-diode operation in the nearly linear range is specified with the aid of a single chart which applies to all materials and to arbitrary values of series resistances. A set of semicircles represents the power–inductance product per unit voltage-swing squared. Another set of quarter circles on the same chart specify the capacitance–inductance products. The general chart is rather simple to construct.

Résumé—On démontre la dérivation d'une expression pour la puissance d'une diode tunnel quasi-linéaire. Cette expression est basée sur le fait que la constante de temps *RC* des jonctions tunnel est une fonction sensible de la concentration d'impuretés dans le matériau du semiconducteur, tandis que le produit résistance–capacité en série est une fonction plus indépendante de la concentration d'impuretés. L'opération de la diode tunnel dans la gamme quasi-linéaire est spécifiée à l'aide d'un seul graphique qui s'applique à tous les materiaux et aux valeurs arbitraires de la résistance en série. Un groupe de demicercles représente le produit puissance–inductance par unité d'oscillation de tension au carré. Un autre groupe de quarts de cercle sur le même graphique spécifie les produits capacité–inductance. Le graphique complet est relativement simple à produire.

Zusammenfassung—Ein Ausdruck für eine fast lineare Leistung einer Tunneldiode wird abgeleitet. Der Ausdruck basiert auf der Tatsache, dass die Zeitkonstante *RC* von Tunnelübergängen eine empfindliche Funktion der Störstellenkonzentration im Halbleitermaterial ist, wogegen das Produkt aus Reihenwiderstand und Kapazität mit dieser nur sehr langsam variiert. Das Arbeiten der Tunneldiode in dem fast linearen Bereich wird durch eine bildliche Darstellung illustriert, die auf alle Materialien und auf willkürliche Werte der Reihenwiderstände anwendbar ist. Eine Gruppe von Halbkreisen stellt das Produkt aus Leistung und Induktion für das Quadrat einer Spannungsschwingung dar. Eine Gruppe von Viertelkreisen stellt das Produkt aus Kapazität und Induktanz dar. Diese bildliche Darstellung lässt sich einfach konstruieren.

I. INTRODUCTION

EXPRESSIONS for tunnel-diode oscillation power for the nearly linear range have been developed recently,[1] and charts specifying diode parameters such as capacitance inductance as a function of power have been given. In Appendix III of the same reference, a general equation for power that omitted the approximation $|R| \gg R_s$ made in the text was also given. R and R_s were respectively the negative and series resistances of the diode.

In this paper, this expression for power output is derived in a simpler manner, and the final results are made more general. They apply to any tunnel diode material. Similarly, the effect of the diode series resistance is eliminated from the plots. Estimates of the parameters are given in one general chart.

II. THEORY

In this Section we shall discuss the class of linear or sinusoidal oscillations in tunnel diodes. The conditions that are necessary for such oscillations and their significance in practical applications will be discussed in Section III. Quantitative estimates of power output will be given in Section IV where a comparison with experimental results will be presented.

The power output of a tunnel diode oscillating sinusoidally at a given angular frequency ω can be estimated with the aid of the linear equivalent

network shown in Fig. 1. R_L and L are the series equivalent load and inductance seen by the diode; the diode is represented by the series resistance R_s and the parallel combination of the negative resistance $-R$ and junction capacitance C. The current I is the r.m.s. value of the current that flows through R_L.

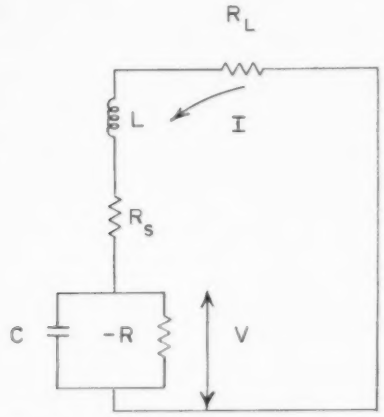


FIG. 1. Equivalent network for tunnel diode. The junction parameters are C, $-R$ and R_s. The inductance of the cavity and the load are L and R_L.

The series impedance of this network is

$$Z = R_L + R_s - \frac{R}{1+R^2C^2\omega^2} - j\frac{R^2C\omega}{1+R^2C^2\omega^2} + j\omega L \tag{1}$$

From Fig. 1, the loop equation under oscillating conditions

$$IZ = 0 \tag{2}$$

must hold. Since the current is finite, sinusoidal steady-state oscillations can occur only when Z is zero. Setting both the real and imaginary parts of Z to zero and solving for the corresponding frequencies leads to

$$\omega^2 = \frac{1}{C^2R^2}\left(\frac{R}{R_L+R_s} - 1\right) \tag{3}$$

in the case of the real part, and to

$$\omega^2 = \frac{1}{C^2R^2}\left(\frac{R^2C}{L} - 1\right) \tag{4}$$

in the case of the reactive part. The power dissipated in the load is

$$P = |I|^2R_L \tag{5}$$

where the square of the current is given by:

$$|I|^2 = \frac{V^2}{8}\left(\frac{1}{R^2} + \omega^2C^2\right) \tag{6}$$

The voltage swing V for the substantially linear part of the voltage-current curve (see Fig. 2) is an inherent property of the junction. It is related to the energy-gap and Fermi levels on the two sides of the junction in the semiconductor, and does not depend on the area of the junction.

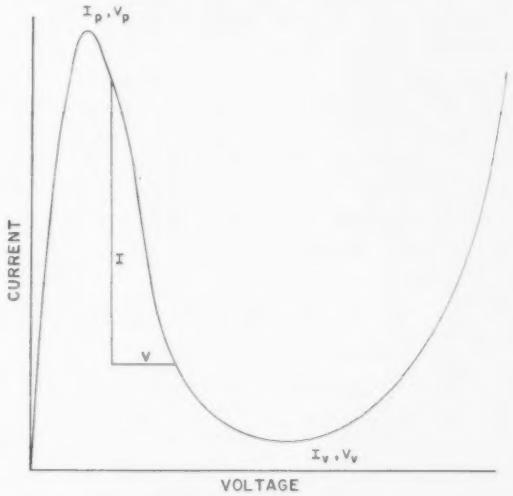


FIG. 2. Typical tunnel-diode current–voltage characteristics with forward bias: V_p peak voltage, V_v valley voltage, V and I voltage and current swings over linear range of negative resistance.

This is in contrast with the peak current I_p, which is directly proportional to the area of the junction and which depends very much upon the semiconductor material. The quantity V depends on doping, but this dependence is very small, because the peak voltages of the diode change only a small amount with doping.

Equations (3)–(6) must all be satisfied simultaneously in order for (5) to describe power for sinusoidal oscillations. Using (3) and (6) we eliminate from (5) the current $|I|^2$ and the load

R_L. After multiplication with C in order to form (RC) in place of R we obtain

$$P = \frac{V^2}{8} C\left(\frac{1+(RC)^2\omega^2}{(RC)^2}\right)\left(\frac{RC}{1+\omega^2(RC)^2} - R_s C\right) \quad (7)$$

Our immediate objective is to obtain an expression that will yield information about the power capability of tunnel diodes in the nearly linear range of oscillation. In constructing tunnel junctions, there are essentially two parameters at one's disposal: the area of the junction, which can be controlled within wide limits, and the resistivity of the material which can be varied by a factor of 2–3. The two parameters can be manipulated independently of each other, and it is desirable to express (7) in terms of one variable, preferably one that is a sensitive function of resistivity. The variable R is a function of area as well as resistivity, but the time constant RC has the desired sensitive dependence on doping, and it is also independent of the area. Taking as variable RC rather than R, we solve (4) for C, and only as a function of ω, L and RC. Substituting this C in (7), we have

$$P = \frac{V^2}{8L}\left(\frac{RC}{1+\omega^2(RC)^2} - R_s C\right) \quad (8)$$

Any power expression for tunnel diodes must give zero power when $R = R_s$. It must also yield $\omega = 0$ for $R = R_s$. It is seen that (8) satisfies these requirements; these are general and must hold irrespective of the type of impedance outside the junction. Since the C, $-R$ and R_s representation of the tunnel junction is generally valid (see Appendix) it is clear from (1) that no oscillation is possible when $R = R_s$. This is true irrespective of any values that one might assign to C, L and R_L.

Before using (8) to estimate power at a given frequency, it is necessary to discuss some of the parameters. RC is a variable which is independent of R_sC.

This holds quite generally because, in the first place, RC is strictly independent of the junction area, and therefore even when R_sC depends on the junction diameter, which is the case with point contact or spreading resistance-type R_s values, RC cannot be made to depend on R_sC by changing the area. Secondly, R_sC turns out to be a very slowly varying function of the impurity concentrations N. According to Shockley's p–n-junction theory for abrupt junctions $C \sim \sqrt{N}$, and, on the other hand, actual measurements of resistivity on As-doped Ge indicate that, due to the severe drop in mobility, the resistivity can be lowered from 0·0008 Ω to 0·0004 Ω-cm only if N is increased[2] by a factor of close to 4. Consequently, in this resistivity range, which is directly applicable to tunnel diodes, R_sC will not change much with doping. As we noted before, RC is a very sensitive function of N, and it is quite generally permissible to change RC independently of R_sC.

At a given frequency, (8) yields a maximum for P when

$$\omega = \frac{1}{RC} \quad (9)$$

which implies that there is an optimum time constant that should be used in order to obtain maximum power output at a given frequency. If higher time constants are used the power will decrease. The maximum power is given by

$$P_{\max} = \frac{V^2}{8L}\left(\frac{1}{2\omega} - R_s C\right) \quad (10)$$

Substituting (9) into (4) we obtain a relationship for the capacitance

$$C = \frac{1}{2L\omega^2} = \frac{2L}{R^2} \quad (11)$$

that leads to the $P_{\max}$ at the operating frequency. The time constant RC is a direct function of the peak-current density of the junction, which can be expressed approximately in terms of the measurable value $I_p/C = K$, giving

$$RC \simeq VC/I_p = V/K \quad (12)$$

Using (9)–(12) we can calculate all the pertinent parameters that optimize the power output.

III. DISCUSSION OF LINEARITY

Since tunnel diodes are strongly nonlinear devices, the assumption of purely sinusoidal oscillations or the linear treatment of the problem can be questioned. Quite generally, the amplitude of oscillation in a tunnel diode is dependent upon

the load resistance. Diodes oscillating close to their resistive cutoff frequency carry out very small amplitude oscillations and only with load resistances approaching zero. On the other hand, if a diode is operated at a frequency considerably below its cutoff and with a very small load resistance, the oscillations become strongly nonlinear. In this case, as the load is increased, the amplitude of the oscillation becomes smaller and approaches zero amplitude[3] as R_L approaches $-R_s+L/RC$. Thus, the amplitude of oscillation is a monotonically varying function of R_L, and generally, by selecting the proper R_L it is possible to cause the tunnel diode to oscillate with voltage amplitudes ranging from zero and up. It is possible, therefore, to consider a class of oscillations which are confined to arbitrary voltage swings V (see Fig. 2) that are selected small enough to limit the excursions along the negative resistance characteristic, to regions as linear as desired.

The next question is whether such a V would lead to only almost zero-power, "threshold" type oscillations. It is evident that one can have as much power as desired in this nearly linear region of V by increasing the area of the junction. Such an increase does not affect the V_v–V_p voltage range of the active region, but increases the current swing corresponding to V. Increasing the area makes the active region contained within V become more linear while at the same time increasing the power available in the region. Thus, in principle it is possible to have oscillations as linear as desired and with as much power as desired. Of course, in actual practice in order to achieve any appreciable amounts of power it is necessary to use extremely low inductances requiring special cavities or resonators.

Examples of power output calculations are given in the next Section. The maximum amount of nearly linear power that a tunnel diode would be expected to handle is of particular interest in tunnel-diode amplifiers where nonlinear distortions can be tolerated only up to a rather limited degree.

In summary, for the linear approach adopted in this paper, the oscillations are confined to a predetermined voltage swing V by using always the proper R_L value that gives such a voltage swing. The method of choosing a practical value of V is discussed in the next Section.

IV. CALCULATION OF EXAMPLES

Assuming $R_sC = 6\ \Omega$-$\mu\mu$F, which is a value that has been attained in practice, and $V = 0{\cdot}1$ and $0{\cdot}15$ V for Ge and GaAs diodes respectively, we calculate the inductances, capacitances and peak currents of junctions that will give the powers between 100 μW and 4 mW at 10, 6 and 1 kMc/s respectively. These results are summarized in Table 1.

The results shown on Table 1 are quite conservative since the V's have been restricted to rather small values. Within the negative-resistance characteristic, these V values correspond to regions in which the average value of V/I in Fig. 2 does not exceed the minimum value of the negative resistance by more than 15 per cent. Under d.c. conditions, the power available within the V region is about 40 per cent of the total power available in the entire negative-resistance region including the strongly nonlinear parts. If it is possible to take advantage of the nonlinear parts as well, these estimated power values could possibly be increased by a factor of about 2. The table also shows the unusually low inductance values that are necessary in order to achieve the output powers indicated.

Experimental values for power output have been published recently.[4] For a GaAs tunnel diode with the following parameters

$$R = 3{\cdot}8\ \Omega \qquad C = 15\ \mu\mu F$$
$$R_s = 1{\cdot}2\ \Omega \qquad L = 0{\cdot}5\ \text{m}\mu\text{H}$$

the power output at 900 Mc/s has been found to be 1·7 mW. If the values of the parameters and the frequency are inserted in (8), we obtain a power of 2·2 mW under the assumption of $V = 0{\cdot}15$ V for GaAs. The results indicate that this diode was probably oscillating with a smaller voltage swing of $V = 0{\cdot}132$ V. This rather satisfactory comparison with experiment appears to point out that the nearly linear region of about 15 per cent variation in R incorporates a substantial portion of the practical power-output capability of a tunnel diode.

V. GENERALIZED TUNNEL-DIODE CHART

Equation (8) can be transformed into such a form that the effects of the series resistance through R_sC, and the effects of V which depend on the material, are no longer present explicitly

in the formula. Defining a new variable

$$M = \frac{8PL}{V^2} + R_s C \tag{13}$$

(8) becomes

$$M = \frac{RC}{1+\omega^2 R^2 C^2} \tag{14}$$

We can now plot a set of constant-M curves with ω as the abscissa and RC the ordinate. However, all points of these curves have to be computed individually. This computation can be avoided by using $1/RC$ instead of RC as the ordinate. If RC is expressed in $\Omega \times \mu\mu F$, $\omega/2\pi$ in kMc/s, and M in $\Omega \times \mu\mu F$ which is also equivalent to $mW \times m\mu H/V^2$, we have

$$\frac{2\pi}{10^3} M = \frac{10^3/2\pi RC}{(10^3/2\pi RC)^2 + f^2} \tag{15}$$

This form of (8) is extremely easy to plot. If we use the frequency f in kMc/s as the abscissa, the inverse time constant $(10^3/2\pi RC)$ as the ordinate, and $2\pi \times 10^{-3} M$ as the arbitrary parameter, we obtain a series of semicircles, each corresponding to a given value of M. All semicircles pass through the origin, have their centers on the axis of ordinates, and for each M value there is an optimum time constant $(RC)_0$ corresponding to the frequency

$$f_0 = 10^3/2\pi(RC)_0 \tag{16}$$

or

$$(2\pi \times 10^{-3} M)_{max} = \frac{1}{2f_0} \tag{17}$$

At a given frequency of operation f_0, (17) gives the maximum attainable value for M. These maxima fall on a straight line on the chart shown in Fig. 3. The M-value of any semicircle is simply the inverse of twice the peak frequency corresponding to that semicircle given by (17). Thus the labeling of the semicircles can be done merely by inspection. The smallest possible M-value is attained when $R_s C = M$, and at this point the semicircle corresponds to the zero-power, resistive cutoff curve.

The chart also shows a series of quarter-circles that correspond to constant capacitance–inductance product curves. The inverse of the radius of each quarter-circle is $\sqrt{(4\pi^2 10^{-3} LC)}$. This can be seen

Table 1

F (*kMc/s*)	RC (*sec*)	K ($mA/\mu\mu F$)	$(PL)_{max}$ (mW–$m\mu H$)	L ($m\mu H$)	P (mW)	C ($\mu\mu F$)	I_p (mA)
Germanium ($V = 0{\cdot}1$V, $R_s C = 6\Omega$–$\mu\mu$F)							
10	16×10^{-12}	6·25	$0{\cdot}25\times10^{-2}$	0·025	0·1	5·1	31·6
6	26×10^{-12}	3·75	$0{\cdot}875\times10^{-2}$	0·01	0·875	35·3	133
1	160×10^{-12}	0·625	$9{\cdot}25\times10^{-2}$	0·025	3·7	510	316
Gallium arsenide ($V = 0{\cdot}15$V, $R_s C = 6\Omega$–$\mu\mu$F)							
10	16×10^{-12}	9·4	$0{\cdot}56\times10^{-2}$	0·056	0·1	2·3	21·3
6	26×10^{-12}	5·6	$1{\cdot}97\times10^{-2}$	0·01	1·97	35·3	197
1	160×10^{-12}	0·94	$20{\cdot}8\times10^{-2}$	0·056	3·7	230	213

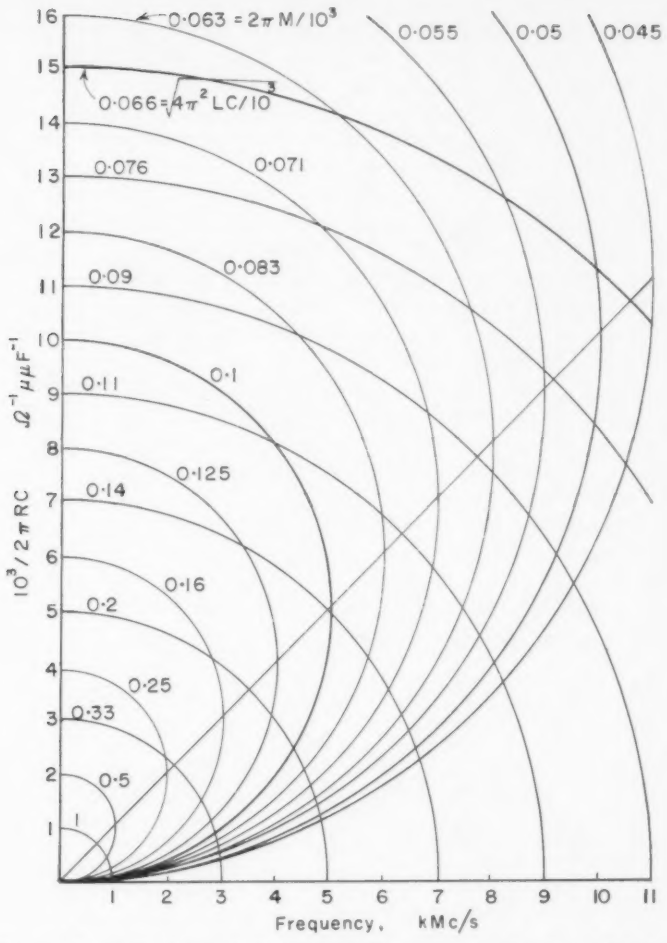


FIG. 3. Generalized tunnel-diode chart. The parameter for the semicircles is $2\pi\times10^{-3}M$ where M is in $\Omega\times\mu\mu$F, or mW×mμH/V². The parameter for the quarter circles is $4\pi^2\times10^{-3}LC$, where L is in mμH and C in μμF. The straight line is the locus of the maximum frequency points of the semicircles.

by writing (4) in the form

$$4\pi^2\times10^{-3}LC=\frac{1}{(10^3/2\pi RC)^2+f^2} \qquad (18)$$

where L is in mμH and C in μμF; f in kMc/s and R in Ω.

VI. CONCLUSIONS

The time constant RC of a tunnel junction is a quantity that is fairly independent of the R_sC product. RC is a very sensitive exponential function of doping, whereas R_sC turns out to be a very slowly varying function of doping. The linear tunnel-diode power analysis developed in this paper rests on this fact.

The major results are as follows.

At a given frequency of operation, only junctions having an optimum $RC=\omega^{-1}$ will produce maximum nearly linear power.

The inductance and the power appear as a product, and by lowering the inductance it is possible to achieve higher power.

The power is proportional to the square of the voltage swing of the oscillations. Actually power, inductance and voltage are expressed as a single variable giving a power-inductance product per unit voltage-swing squared.

The effect of R_s is to reduce the available power in the load. This reduction is very considerable when R_sC approaches RC, but it is negligible when R_sC is about twenty times smaller than RC.

The predictions concerning the inductance and power values appear to be in agreement with certain experimental results. Thus for some purposes a linear analysis of tunnel diodes appears to give fairly satisfactory results.

A wide range of tunnel-diode sinusoidal oscillation is specified by a single chart, where a set of simple semicircles for constant M-values gives the power–inductance products per unit voltage-swing squared, and a set of quarter-circles gives the capacitance–inductance products as a function of frequency and inverse time constant.

The chart is constructed in a simple way, and is independent of the effects of series resistance.

Acknowledgements—The author is indebted to Messrs. W. HAUER, E. MURPHY, B. TYSON and H. LOCKWOOD, and Dr. S. MAYBURG and Dr. P. H. KECK whose stimulating discussions made this paper possible.

APPENDIX

Tunnel-diode junctions for most applications must be made very small in order to obtain high enough impedance levels in the operating circuit. The low time constants necessary for high-frequency operation can in practice be used only if the junction is made small. This fact gives validity to the simple equivalent representation of the junction, since the skin effects turn out to be negligible up to rather high frequencies. The fact that the resistivity of the semiconductor ohmic regions are high, at least ten times higher than the resistivity of copper, is an additional reason for neglecting skin effects.

The remaining elements of the equivalent circuit in Fig. 1, are the impedance L and load R_L seen by the junction. We are justified in not including any series capacitance in this external circuit, due to the fact that the condition for high-frequency oscillation requires very small inductances which means small physical dimensions, resulting in a negligible capacitance compared to the junction capacitance.

Quantitative estimates for the limits of the idealized circuit were discussed in Ref. 1.

REFERENCES

1. G. DERMIT, *Proc. I.R.E.* **49,** 1033 (1961). (General references to tunnel diode literature are given in this paper.)
2. R. N. HALL, *Application of Tunneling to Active Diodes*. Scientific Report No. 3A, Contract USAF 19(604)–6623 (1960).
3. M. E. HINES, *Bell Syst. Tech. J.* **39,** 477 (1960).
4. R. STERZER and D. NELSON, *Proc. I.R.E.* **49,** 744 (1961).

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 215–225. Printed in Great Britain

THEORETICAL CURRENT GAIN OF A CYLINDRICAL MESA TRANSISTOR

DAVID P. KENNEDY and PHILIP C. MURLEY

Product Development Laboratory, Data Systems Division, International Business Machines Corporation, Poughkeepsie, New York

(*Received* 13 *April* 1961)

Abstract—For the cylindrical mesa transistor, computation of base-region transport efficiency is considered a boundary-value problem; solution of this problem yields mathematical equations applicable to the design and development of practical semiconductor devices. This analytical method is applied to the problem of minority-carrier transport across a solid-cylinder type structure which approximates the transistor base region. Minority-carrier losses—representing a fundamental limitation upon transistor-current gain—are introduced through an assumption of bulk and surface-recombination mechanisms.

Applications of this analysis are illustrated by establishing the common-emitter current gain for typical junction transistors. Assuming, in such computations, physical parameters approximating germanium and silicon diffused devices, the silicon transistor is shown to be less sensitive to surface recombination mechanisms. Further, the existence of an optimum emitter radius is demonstrated for a semiconductor structure similar to the cylindrical hook collector.

Résumé—Pour le transistor mesa cylindrique, le calcul du rendement de transport de la région de base est effectué en considérant le problème des veleurs de bornes données; la solution de ce problème produit des équations différentielles applicables au dimensionnement et au developpement de dispositifs de semiconducteurs pratiques. Cette méthode analytique est appliquée au probléme du transport de porteurs minoritaires à travers une structure de type cylindrique solide ressemblant à la region de base du transistor. Les pertes de porteurs minoritaires—représentant une limitation fondamentale au gain de courant du transistor—sont introduites en assumant des mécanismes de recombinaison totale de surface.

Des applications de cette analyse sont illustrées en établissant le gain commun du courant emetteur pour les transistors de jonctions typiques. En assumant dans ces calculs des paramètres physiques ressemblant aux elements diffusés au germanium et au silicium, on démontre que le transistor au silicium est moins sensible aux mécanismes de recombinaison de surface. En plus, on démontre l'existence d'un rayon emetteur optimum pour une structure de semiconducteur similaire au crochet cylindrique du collecteur.

Zusammenfassung—Für den zylindrischen Mesa-Transistor wird die Berechnung der Transportleistung in der Basis als ein Randwertproblem aufgefasst. Die Lösung des Problems ergibt mathematische Gleichungen, die auf den Bau and die Entwicklung von Halbleitergeräten anwendbar sind. Dieses analytische Verfahren wird auf das Problem des Transports von Minoritätsträgern durch eine feste zylindrische Struktur angewendet, die der Basiszone des Transistors entspricht. Verluste an Minoritätsträgern, die eine fundamentale Beschränkung der Stromverstärkung des Transistors darstellen, werden durch die Annahme eines Rekombinationsmechanismus im Gesamtkörper und an der Oberfläche erklärt.

Für typische Übergangstransistoren wird die allgemeine Emitterstrom–Verstärkung auf Grund dieser Analyse berechnet. Dabei werden physikalische Parameter eingeführt, die näherungsweise denen von diffundierten Silizium- und Germanium-Geräten entsprechen. Es ergibt sich, dass der Silizium-Transistor für die Oberflächenrekombination weniger empfindlich ist. Ausserdem wird nachgewiesen, dass für eine Halbleiterstruktur, die dem zylindrischen hook-Kollektor ähnlich ist, ein optimaler Emitterradius existiert.

I. INTRODUCTION

CONTROL of common-emitter current gain represents an important problem in the design and development of transistors. Two apparently identical devices can—for no obvious reason—exhibit significantly different magnitudes of current gain; these differences are often attributed to base-region transport efficiency which is modified by surface-recombination mechanisms. This paper presents an analytical investigation into the minority-carrier transport problems as encountered in a cylindrical mesa type of drift transistor, Fig. 1. Computation of the base-region minority-carrier distribution is considered a boundary-value problem having mixed boundary conditions;[1] this situation results from the presence of an emitter junction, an ohmic contact, and exposed semiconductor material—all at one end of the structure. The solution of this boundary problem yields a mathematical expression for the common-emitter current gain, due to base-region transport, in terms of the physical and geometrical parameters of a practical transistor.

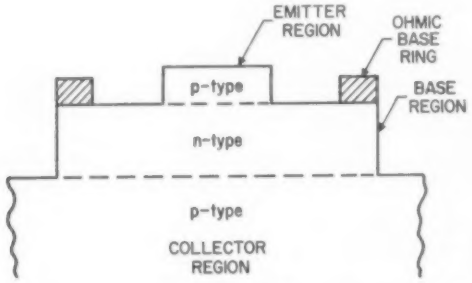


FIG. 1. Cross section of cylindrical mesa transistor.

For analytical purposes, the cylindrical transistor base region, Fig. 1, is approximated by a solid circular cylinder of n-type semiconductor material ($0 < r < c$; $0 < z < w$) with a circularly symmetrical emitter junction at one end ($0 < r < a$; $z = 0$) and a collector junction at the other ($0 < r < c$; $z = w$), Fig. 2. The emitter junction is assumed to inject a constant-density minority-carrier current, J_{pe}, into this structure—resulting excess minority carriers are subject to both bulk and surface-recombination mechanisms, thereby limiting the base-region transport efficiency. An axially directed drift field of constant magnitude E_z is used to characterize the influence of conductivity grading within diffused transistors; this type of field distribution could result from an exponential impurity distribution between the emitter and collector junctions.

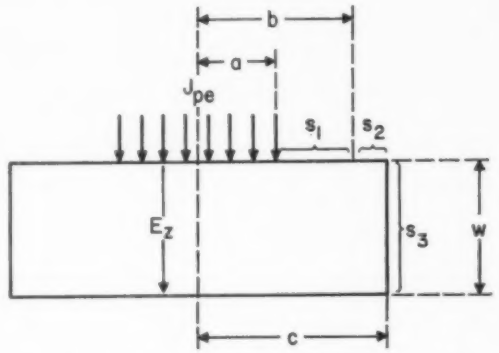


FIG. 2. Analytical model of cylindrical base region.

II. NOTATION

a	emitter radius
b	inner radius of base ring
c	mesa radius
k	Boltzmann constant
q	electron charge
r	radial co-ordinate
s	surface-recombination velocity
w	base width
z	axial co-ordinate
D_p	hole-diffusion constant
L_p	hole-diffusion length
J_p	hole current density
T	temperature (absolute)
τ_p	hole lifetime
λ	arbitrary separation constant

III. ANALYSIS

The steady-state distribution of excess minority carriers (holes) within this cylindrical base region, $P(r; z)$, is established by the continuity equation

$$\frac{\partial^2 P}{\partial r^2} + \frac{1}{r}\frac{\partial P}{\partial r} + \frac{\partial^2 P}{\partial z^2} - f\frac{\partial P}{\partial z} - \frac{P(r,z)}{L_p^2} = 0 \tag{1}$$

$$f = \frac{qEz}{kT}$$

Assuming the minority-carrier distribution can be represented by

$$P(r;z) = X(r)Y(z) \tag{2}$$

this continuity equation is separable into two ordinary differential equations,

$$\frac{d^2x}{dr^2} + \frac{1}{r}\frac{dx}{dr} + \lambda^2 X(r) = 0 \tag{3a}$$

$$\frac{d^2y}{dz^2} - f\frac{dy}{dz} - (\lambda^2 + L_p^{-2}) = 0 \tag{3b}$$

It is important to note that the nature of this problem restricts solutions of equation (3a) to zero-order Bessel functions. The minority-carrier distribution $P(r;z)$ must be well behaved at the origin ($r = 0$); therefore Neumann functions are excluded from this solution. From equations (2) and (3) we therefore have the general distribution equation

$$P(r;z) = e^{fz/2}\{A\cosh((\gamma z) + B\sinh(\gamma z)\}J_0(\lambda r) \tag{4}$$

where

$$\gamma^2 = \lambda^2 + L_p^{-2} + (f/2)^2 \tag{5}$$

The arbitrary constants associated with equation (4) must be adjusted to satisfy all boundary conditions imposed upon this solid cylinder by practical considerations. Defining surface-recombination velocity in the usual manner,[2] the transistor base region is characterized by a combination of Dirichlet and Cauchy type boundary equations[3] which are given by

$$\frac{\partial P}{\partial r} + \frac{S_3}{D_p}P(r;z) = 0 \qquad r = c; 0 < z < w \tag{6a}$$

$$P(r;z) = 0 \qquad 0 < r < c; z = w \tag{6b}$$

$$\frac{\partial P}{\partial r} - fP(r;z) = \begin{cases} -\dfrac{J_{pe}}{qD_p} & 0 < r < a; z = 0 \\ \dfrac{S_1}{D_p}P(r;z) & a < r < b; z = 0 \\ \dfrac{S_2}{D_p}P(r;z) & b < r < c; z = 0 \end{cases} \tag{6c}$$

The outer wall of this solid cylinder ($r = c$; $0 < z < w$) is assumed to exhibit a recombination current proportional to an arbitrary surface-recombination velocity S_3 and the concentration of excess minority carriers at this location, equation (6a). The collector junction, which is an ideal minority-carrier sink, has been represented by a surface that is completely free of excess minority carriers, equation (6b). The entire emitter end of this solid cylinder ($0 < r < c$; $z = 0$) is characterized by equation (6c). A constant minority-carrier current density J_p is assumed to result from a symmetrically located, forward-biased emitter junction ($0 < r < a$; $z = 0$); the remainder of this surface exhibits recombination mechanisms due to an exposed semiconductor surface ($a < r < b$; $z = 0$) and the metal–semiconductor interface of a base-region ohmic contact ($b < r < c$; $z = 0$).

The computational difficulties inherent in this problem have required an introduction of simplifying assumptions for the emitter-junction boundary conditions. Characterization of an emitter junction as a current source—rather than a source of minority carriers—eliminates an explicit computation of total emitter current; an analytical comparison of these boundary conditions has shown no observable difference in transport efficiency when applied to elementary cylindrical structures. Furthermore, the assumption of a constant minority-carrier current density at this location is found only in transistors exhibiting a negligible radial biasing potential at the emitter junction. Because this situation results from an electron current immediately under the emitter junction, this investigation is restricted to the low-current operation of transistors exhibiting small magnitudes of bulk minority-carrier recombination and emitter-junction electron current.

A negligible limitation should result from the assumed constant magnitude minority-carrier drift field. Previous investigations on this subject[4] indicate the drift field in a diffused-type transistor can be approximated by one of constant magnitude; this constant is obtained at the base-region edge of the emitter-junction depletion layer.

Substituting equation (4) into equation (6a)

yields the relation

$$P(r;z) = e^{fz/2}\sum_{a=1}^{\infty}\left\{A_a \cosh\left(\delta_a\frac{z}{c}\right) + B_a \sinh\left(\delta_a\frac{z}{c}\right)\right\}J_0\left(\alpha_a\frac{r}{c}\right) \tag{7}$$

where

$$\delta_a^2 = \alpha_a^2 + \left(\frac{fc}{2}\right)^2 + \left(\frac{c}{L_p}\right)^2 \tag{8a}$$

$$0 = \alpha_a J_1(\alpha_a) - \frac{S_3 c}{D_p}J_0(\alpha_a) \tag{8b}$$

The summation of equation (7) is conducted over ascending roots of equation (8b)—there are an infinite number of such roots for each assumed value of surface-recombination velocity. Further, substituting for equations (6c) the conditional boundary equation

$$\frac{\partial P}{\partial z} - fP(r;z) = -\sum_{a=1}^{\infty}\psi_a J_0\left(\alpha_a\frac{r}{c}\right) \quad 0<r<c;\, z=0 \tag{9}$$

we have, for equation (7)

$$P(r;z) = \exp(fz/2)\sum_{a=1}^{\infty}\psi_a \times \psi_a\left\{\frac{c\sinh[(\delta_a/c)(w-z)]}{\frac{fc}{2}\sinh\left(\delta_a\frac{w}{c}\right)+\delta_a\cosh\left(\delta_a\frac{w}{c}\right)}\right\}J_0\left(\alpha_a\frac{r}{c}\right) \tag{10}$$

Equation (10) satisfies all boundary equations of this problem except at the emitter end of the base region; at this location we have, from equations (6c) and equation (9),

$$\sum_{a=1}^{\infty}\psi_a J_0\left(\alpha_a\frac{r}{c}\right) = \begin{cases}\dfrac{J_{pe}}{qD_p} & 0<r<a \\ -\dfrac{S_1}{D_p}P(r;z) & a<r<b \\ -\dfrac{S_2}{D_p}P(r;z) & b<r<c\end{cases} \tag{11}$$

Equation (11), in this form, is exceedingly difficult to solve and therefore has been modified by assuming that the total minority-carrier distribution, $P(r;z)$, is the sum of three distribution equations which are essentially the same as equation (10),

$$P(r;z) = P_1(r;z) + P_2(r;z) + P_3(r;z) \tag{12}$$

Using this technique it is shown (see Appendix) that the base-region minority carrier distribution is given by

$$P(r;z) = \frac{2J_{pe}a}{qD_p}\exp(fz/2)\sum_{n=1}^{\infty}E_n \times \left\{\frac{\sinh[(\delta_n/c)(w-z)]}{\frac{fc}{2}\sinh\left(\delta_n\frac{w}{c}\right)+\delta_n\cosh\left(\delta_n\frac{w}{c}\right)}\right\}J_0\left(\alpha_n\frac{r}{c}\right) \tag{13}$$

where

$$\frac{2J_{pe}a}{qD_p c}[E_n] = [\psi_m] + [\Phi_k] + [\Omega_l] \tag{14}$$

and

$$\begin{bmatrix}\psi_m \\ \Phi_k \\ \Omega_l\end{bmatrix} = \frac{2J_{pe}a}{qD_p c}\begin{bmatrix}1 & -J_{m,k} & -K_{m,l} \\ M_{k,m} & 1+\dfrac{S_1}{D_p}A_k & -L_{k,z} \\ Q_{l,m} & N_{l,k} & 1+\dfrac{S_2}{D_p}A_l\end{bmatrix}^{-1} \times \begin{bmatrix}X_m \\ 0 \\ 0\end{bmatrix} \tag{15}$$

The matrix elements of equation (15) are established by equation (A.7).

From equation (13), which fully describes the base-region minority-carrier distribution, we can easily determine the transport efficiency of this structure. Since the axially directed minority-carrier current, $J_{p,z}(r;z)$ is given by

$$J_{p,z}(r;z) = -qD_p\left\{\frac{\partial P}{\partial z} - fP(r;z)\right\} \tag{16}$$

we have, from equation (13),

$$J_{p,z}(r;z) = 2J_{pe}(a/c)\exp(fz/2)\sum_{n=1}^{\infty} E_n \frac{(fc/2)\sinh[(\delta_n/c)(w-z)] + \delta_n \cosh[(\delta_n/c)(w-z)]}{\frac{fc}{2}\sinh\left(\delta_n\frac{w}{c}\right) + \delta_n\cosh\left(\delta_n\frac{w}{c}\right)} J_0[\alpha_n(r/c)] \tag{17}$$

Integrating this expression across the collector junction face ($0 < r < c; z = w$) yields the total collector current I_{pc} and, therefore,

$$B = \frac{I_{pc}}{I_{pe}} = 4\left(\frac{c}{a}\right)\exp(fw/2)\sum_{n=1}^{\infty} \frac{E_n\delta_n J_1(\alpha_n)}{\alpha_n\left[\frac{fc}{2}\sinh\left(\delta_n\frac{w}{c}\right) + \delta_n\cosh\left(\delta_n\frac{w}{c}\right)\right]} \tag{18}$$

From equation (18) we now have

$$\beta_B = \frac{B}{1-B} \tag{19}$$

Equation (19) established the common-emitter current gain resulting from base-region recombination mechanisms;* an emitter injection efficiency γ of unity is assumed. If, in a particular transistor, this assumption cannot be made, the total common-emitter current gain β_T is given by[6]

$$\frac{1}{\beta_T} = \frac{1}{\beta_B} + \frac{1}{\beta\gamma} \tag{20}$$

where

$$\beta_\gamma = \frac{\gamma}{1-\gamma} \tag{21}$$

* Computer programs[5] for equation (19) are available on request to authors.

IV. JUNCTION-TRANSISTOR CURRENT GAIN

Figs. 3 and 4 illustrate the theoretical current gain, equation (19), of two transistors which are geometrically the same; their fundamental difference resides in assumed values of minority-carrier lifetime τ_p and drift mobility μ_p. The first illustration, Fig. 3, establishes the gain of a device exhibiting characteristics of a typical diffused silicon structure ($\tau_p = 10^{-7}$; $\mathbf{D}_p = 6$) when large values of surface-recombination velocity have been assumed ($S_1 = S_3 = 5\times10^3$; $S_2 = 10^4$). Similarly, using the same magnitudes of surface-recombination velocity, the current gain is established for a transistor with base-region properties approximating a germanium transistor ($\tau_p = 2\times10^{-5}$; $D_p = 20$), Fig. 4.

A comparison of Figs. 3 and 4 shows that fundamental differences result from the assumed values of minority-carrier lifetime and drift mobility. The magnitude of current gain presented in Fig. 3 is approximately 20 per cent below the value attained when surface-recombination mechanisms are eliminated—as encountered in a one-dimensional solution of this problem.[7] In contrast, Fig. 4—approximating a typical germanium transistor—exhibits a current gain which is an order of magnitude below that of a one-dimensional structure. From these examples, it appears the decreased minority-carrier diffusion

length of a diffused silicon transistor introduces the need of a narrow base width to obtain useful magnitudes of current gain; this situation reduces surface-recombination current and therefore decreases the influence of such recombination mechanisms.

Eliminating all surface recombination at the exposed base-region surfaces ($S_1 = S_3 = 0$) is insufficient to yield an adequate current gain in the device illustrated by Fig. 4. Minority-carrier losses occurring at the ohmic contact become an important mechanism when a large diffusion length exists in the transistor base region. Fig. 5 illustrates this characteristic for a transistor which has a current gain of approximately 100—assuming $L_p/w = 7{\cdot}0$ and $E_z = 0$—when all surface recombination mechanisms have been eliminated.

An important parameter in the design of junction devices is the emitter-to-base-ring spacing. Advantages obtained through a reduction of this spacing—thereby decreasing extrinsic base resistance—are overcome by an increase of base-contact recombination current and, therefore, a reduction of current gain. Fig. 6 illustrates the relation between current gain β_B and emitter-ohmic contact spacing for a mesa structure of 16-mil radius. As in our previous example ($L_p/w = 7{\cdot}0$; $E_z = 0$), a current gain of approximately 100 is obtained when the ohmic base ring is an adequate distance from the emitter junction.

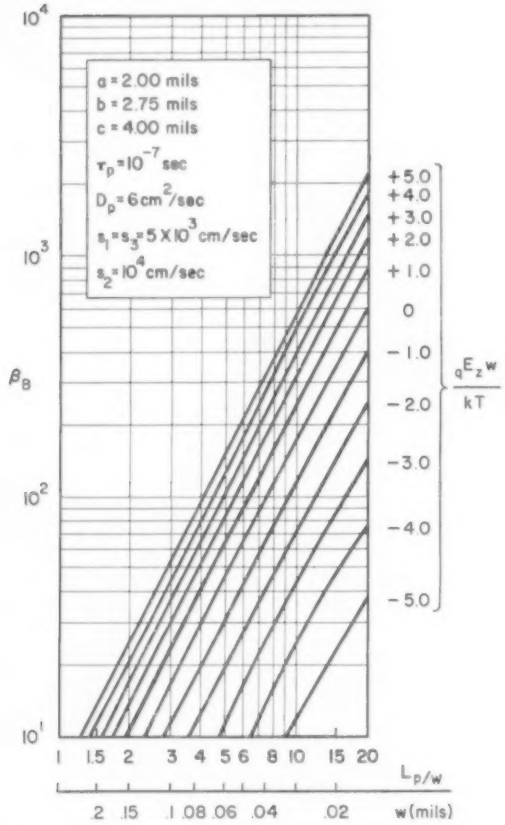


FIG. 3. Current gain of cylindrical mesa transistor—L_p small.

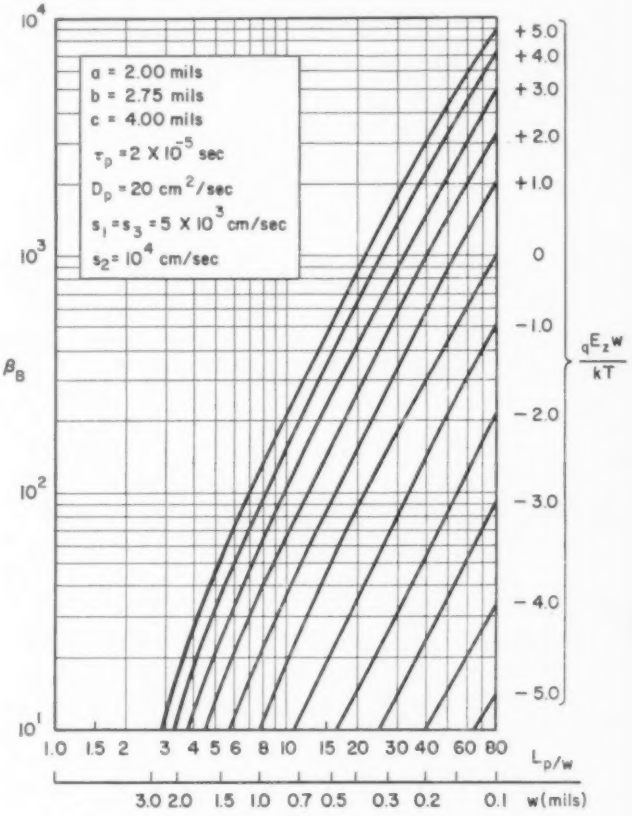


FIG. 4. Current gain of cylindrical mesa transistor—L_p large.

The small sensitivity to surface recombination, illustrated in Fig. 3, results from a substantial bulk recombination current which will, in effect, "swamp out" surface losses. A similar characteristic is obtainable in devices exhibiting a relatively long base-region diffusion length. Fig. 7 presents the influence of surface recombination (S_1; S^6) upon current gain, for two typical transistors—one transistor is assumed to have a large base-region drift field while the other has none. The lower two curves in Fig. 7 characterize these devices throughout a practical range of surface-recombination velocity; the base rings have been adjusted to introduce a negligible minority-carrier loss. Similarly, the upper two curves illustrate the same transistors when a high-recombination ohmic contact is placed near (0·5 mil) the emitter junction; the base width (w) is reduced to restore current gain (at $S_1 = 0$) to its original magnitude. This example indicates that a base contact of controlled recombination velocity can be used to reduce the influence of recombination mechanisms at an exposed base-region surface.

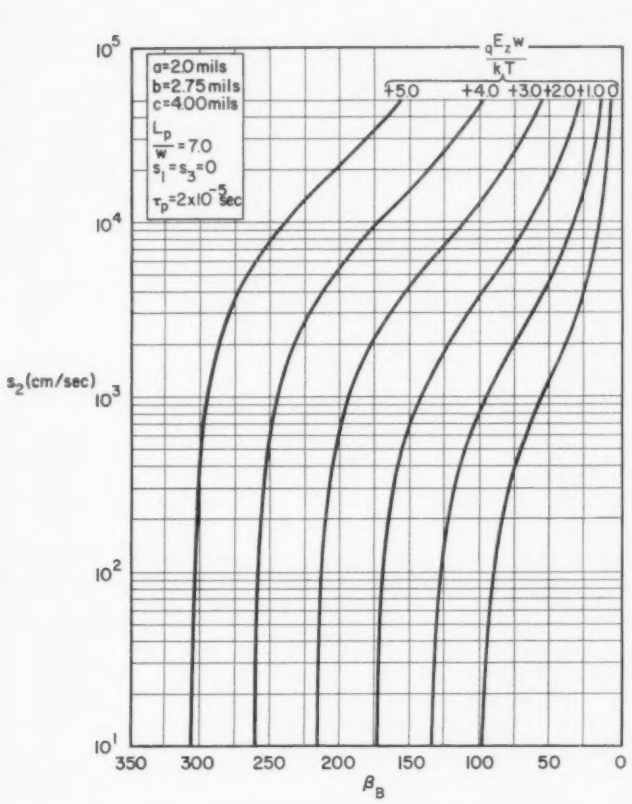


FIG. 5. Current gain vs. ohmic-contact recombination velocity.

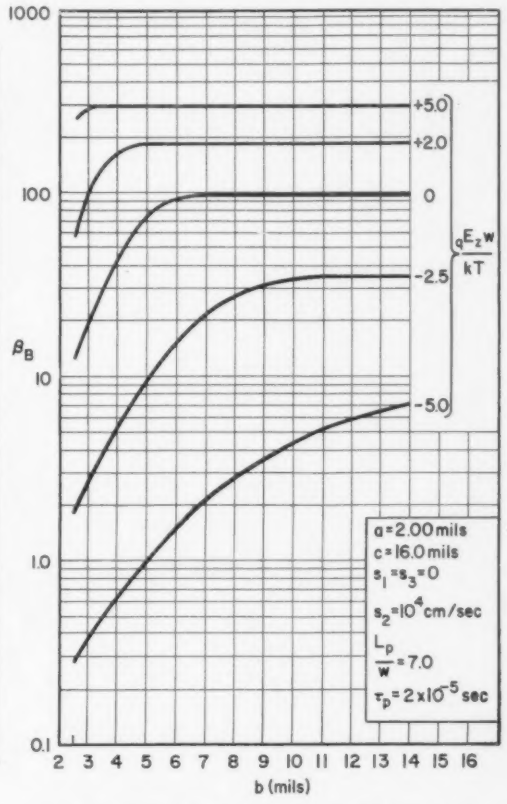


FIG. 6. Current gain vs. ohmic contact–emitter distance.

The last example illustrating applications of equation (19) relates to the current multiplication of a cylindrical hook collector.[6] This structure is essentially identical to Fig. 1 except for the ohmic contact; it is mathematically eliminated by setting $S_2 = 0$ and $b = c$. Current multiplication, being essentially equivalent to the current gain

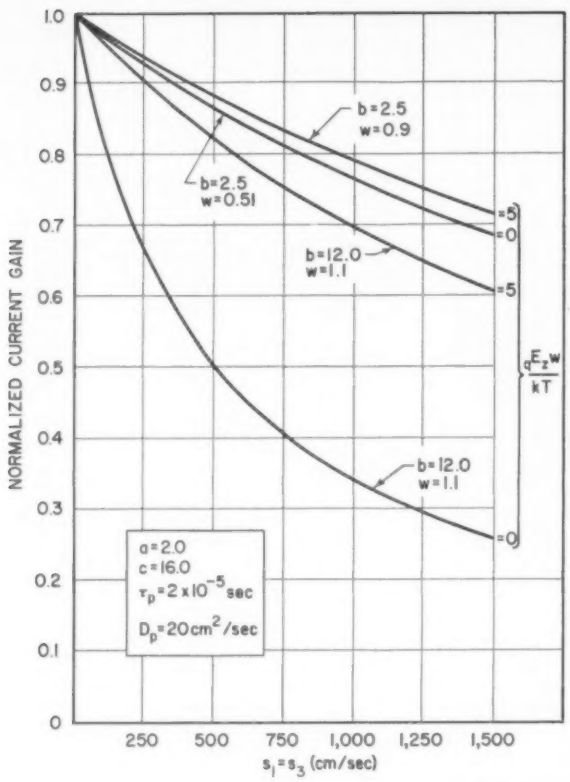


FIG. 7. Current gain vs. surface recombination velocity.

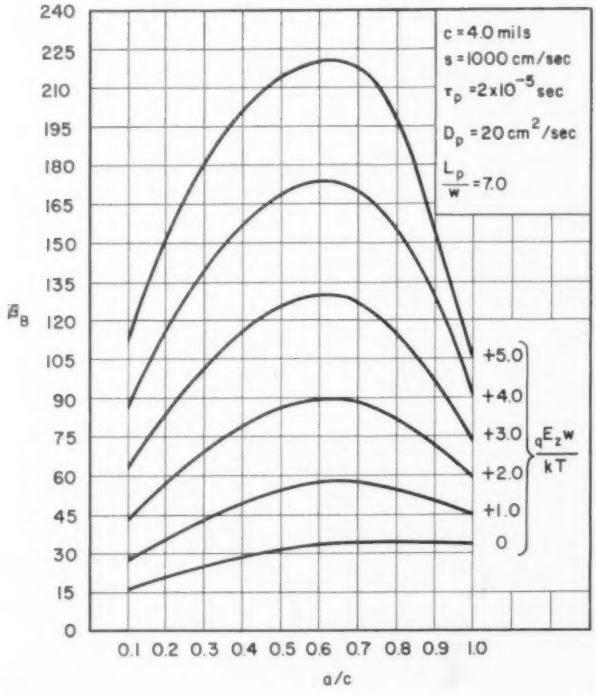


FIG. 8. Current gain vs. emitter radius.

expressed by equation (19), can be maximized by an adjustment of the emitter radius. Fig. 8 illustrates the current gain of a typical structure throughout a range of physical and geometrical parameters. Significant increase in gain is obtainable when the emitter radius has been properly selected; furthermore, at this optimum geometry one obtains a minimum dependence upon emitter radius.

Acknowledgements—The authors wish to thank Mr. W. E. HARDING for his continued support of this project and Mr. A. H. BERGER for many stimulating and helpful discussions.

APPENDIX

The boundary conditions at the emitter end of this base region, equation (6c), have been satisfied by assuming the total minority-carrier distribution, $P(r;z)$, is expressed as the sum of three independent distributions

$$P(r;z) = P_1(r;z)+P_2(r;z)+P_3(r;z) \tag{A.1}$$

where each term is essentially the same as equation (10)

$$P_1(r;z) = \exp(fz/2)\sum_{m=1}^{\infty}\psi_m\left\{\frac{c\sinh[(\delta_m/c)(w-z)]}{\frac{fc}{2}\sinh\left(\delta_m\frac{w}{c}\right)+\delta_m\cosh\left(\delta_m\frac{w}{c}\right)}\right\}J_0[\alpha_m(r/c)] \tag{A.2a}$$

$$P_2(r;z) = \exp(fz/2)\sum_{k=1}^{\infty}\Phi_k\left\{\frac{c\sinh[(\delta_k/c)(w-z)]}{\frac{fc}{2}\sinh\left(\delta_k\frac{w}{c}\right)+\delta_k\cosh\left(\delta_k\frac{w}{c}\right)}\right\}J_0[\alpha_k(r/c)] \tag{A.2b}$$

$$P_3(r;z) = \exp(fz/2)\sum_{l=1}^{\infty}\Omega_l\left\{\frac{c\sinh[(\delta_l/c)(w-z)]}{\frac{fc}{2}\sinh\left(\delta_l\frac{w}{c}\right)+\delta_l\cosh\left(\delta_l\frac{w}{c}\right)}\right\}J_0[\alpha_l(r/c)] \tag{A.2c}$$

Substituting equation (A.1) into the boundary equations (6c), we obtain

$$\sum_{m=1}^{\infty}\psi_m J_0\left(\alpha_m\frac{r}{c}\right) = \frac{J_{pe}}{qD_p}-\sum_{k=1}^{\infty}\Phi_k J_0\left(\alpha_k\frac{r}{c}\right)-\sum_{l=1}^{\infty}\Omega_l J_0\left(\alpha_l\frac{r}{c}\right) \qquad (0<r<a) \tag{A.3a}$$

$$\sum_{k=1}^{\infty}\Phi_k\left(1+\frac{S_1}{D_p}A_k\right)J_0\left(\alpha_k\frac{r}{c}\right) = -\sum_{m=1}^{\infty}\psi_m\left(1+\frac{S_1}{D_p}A_m\right)J_0\left(\alpha_m\frac{r}{c}\right)$$

$$-\sum_{l=1}^{\infty}\Omega_l\left(1+\frac{S_1}{D_p}A_l\right)J_0\left(\alpha_l\frac{r}{c}\right) \qquad (a<r<b) \tag{A.3b}$$

$$\sum_{l=1}^{\infty}\Omega_l\left(1+\frac{S_2}{D_p}A_l\right)J_0\left(\alpha_l\frac{r}{c}\right) = -\sum_{m=1}^{\infty}\psi_m\left(1+\frac{S_2}{D_p}A_m\right)J_0\left(\alpha_m\frac{r}{c}\right)-\sum_{k=1}^{\infty}\Phi_k\left(1+\frac{S_2}{D_p}A_k\right)J_0\left(\alpha_k\frac{r}{c}\right)$$

$$(b<r<c) \tag{A.3c}$$

where

$$A_a = \left[\frac{fc}{2} + \delta_a \operatorname{ctnh}\left(\frac{\delta_a w}{c}\right)\right]^{-1} \tag{A.4a}$$

$$\alpha_a J_1(\alpha_a) - \frac{S_c}{D_p} J_0(\alpha_a) = 0 \tag{A.4b}$$

Equations (A.3) are a Dini–Bessel type of series[7,8] which can be evaluated in a conventional manner. In this evaluation we assume the equality exists only throughout the range specified for each expression; outside this range the left-hand sides of equations (A.3 are assumed to be zero and therefore

$$\sum_{m=1}^{\infty} \psi_m J_0\left(\alpha_m \frac{r}{c}\right) = \frac{J_{pe}}{qD_p} + \frac{S_1}{D_p} \sum_{k=1}^{\infty} \Phi_k A_k J_0\left(\alpha_k \frac{r}{c}\right) + \sum_{l=1}^{\infty} \Omega_l J_0\left(\alpha_l \frac{r}{c}\right) \qquad (0 < r < a) \tag{A.5a}$$

$$\sum_{k=1}^{\infty} \Phi_k \left(1 + \frac{S_1}{D_p} A_k\right) J_0\left(\alpha_k \frac{r}{r}\right) = \left(\frac{S_2 - S_1}{D_p}\right) \sum_{l=1}^{\infty} \Omega_l A_l J_0\left(\alpha_l \frac{r}{c}\right) - \frac{S_1}{D_p} \sum_{m=1}^{\infty} \psi_m A_m J_0\left(\alpha_m \frac{r}{c}\right) \qquad (a < r < b) \tag{A.5b}$$

$$\sum_{l=1}^{\infty} \Omega_l \left(1 + \frac{S_2}{D_p} A_l\right) J_0\left(\alpha_l \frac{r}{c}\right) = -\left(\frac{S_2 - S_1}{D_p}\right) \sum_{k=1}^{\infty} \Phi_k A_k J_0\left(\alpha_k \frac{r}{c}\right) - \frac{S_2}{D_p} \sum_{m=1}^{\infty} \psi_m A_m J_0\left(\alpha_m \frac{r}{c}\right) \qquad (b < r < c) \tag{A.5c}$$

Upon multiplying equation (A.5) by $r J_0[\alpha_j(r/c)]$ and integrating, we obtain

$$\frac{2J_{pe}a}{qD_pc} \begin{bmatrix} X_m \\ \hline 0 \\ \hline 0 \end{bmatrix} = \left[\begin{array}{c|c|c} 1 & -J_{m,k} & -K_{m,} \\ \hline M_{k,m} & 1 + \dfrac{S_1}{D_p} A_k & -L_{k,l} \\ \hline Q_{l,m} & N_{l,k} & 1 + \dfrac{S_2}{D_p} A_l \end{array}\right] \times \begin{bmatrix} \psi_m \\ \hline \Phi_k \\ \hline \Omega_l \end{bmatrix} \tag{A.6}$$

$$J_{m,k} = \left(\frac{2S_1A_k}{D_p}\right) \frac{\int_0^a r J_0\left(\alpha_k \frac{r}{c}\right) J_0\left(\alpha_m \frac{r}{c}\right) dr}{c^2[J_0^2(\alpha_m) + J_1^2(\alpha_m)]} \qquad K_{m,l} = \left(\frac{2S_2A_l}{D_p}\right) \frac{\int_0^a r J_0\left(\alpha_l \frac{r}{c}\right) J_0\left(\alpha_m \frac{r}{c}\right) dr}{c^2[J_0^2(\alpha_m) + J_1^2(\alpha_m)]}$$

$$L_{k,l} = \left(\frac{2(S_2 - S_1)A_l}{D_p}\right) \frac{\int_a^b r J_0\left(\alpha_l \frac{r}{c}\right) J_0\left(\alpha_k \frac{r}{c}\right) dr}{c^2[J_0^2(\alpha_k) + J_1^2(\alpha_k)]} \qquad M_{k,m} = \left(\frac{2S_1A_m}{D_p}\right) \frac{\int_a^b r J_0\left(\alpha_m \frac{r}{c}\right) J_0\left(\alpha_k \frac{r}{c}\right) dr}{c^2[J_0^2(\alpha_k) + J_1^2(\alpha_k)]} \tag{A.7}$$

$$N_{l,k} = \left(\frac{2(S_2-S_1)A_k}{D_p}\right) \frac{\ll \int_b^c rJ_0\left(\alpha_k\frac{r}{c}\right)J_0\left(\alpha_l\frac{r}{c}\right)dr}{c^2[J_0^2(\alpha_l)+J_1^2(\alpha_l)]} \qquad Q_{l,m} = \left(\frac{2S_1A_m}{D_p}\right) \frac{\int_b^c rJ_0\left(\alpha_m\frac{r}{c}\right)J_0\left(\alpha_l\frac{r}{c}\right)dr}{c^2[J_0^2(\alpha_l)+J_1^2(\alpha_l)]}$$

$$X_m = \frac{J_1\left(\alpha_m\frac{a}{c}\right)}{\alpha_m[J_0^2(\alpha_m)+J_1^2(\alpha_m)]}$$

From equation (A.7) we can now establish the terms Ψ_m, Φ_k, and Ω_l, which are required for the solution of this problem,

$$\begin{bmatrix} \psi_m \\ \hline \Phi_k \\ \hline \Omega_l \end{bmatrix} = \frac{2J_{pe^a}}{qD_pc} \left[\begin{array}{c|c|c} 1 & -J_{m,k} & -K_{m,l} \\ \hline M_{k,m} & 1+\dfrac{S_1}{D_p}A_k & -L_{k,l} \\ \hline Q_{l,m} & N_{l,k} & 1+\dfrac{S_2}{D_p}A_l \end{array}\right]^{-1} \times \begin{bmatrix} X_m \\ \hline 0 \\ \hline 1 \end{bmatrix} \tag{A.8}$$

Therefore, upon collecting terms within equation (A.1) we obtain

$$P(r;z) = \frac{2J_{pe^a}}{qD_p}\exp(fz/2)\sum_{n=1}^{\infty} E_n \frac{\sinh[(\delta_n/c)(w-z)]}{\frac{fc}{2}\sinh\left(\delta_n\frac{w}{c}\right)+\delta_n\cosh\left(\delta_n\frac{w}{c}\right)} J_0[\alpha_n(r/c)] \tag{A.9}$$

where

$$\frac{2J_{pe^a}}{qD_pc}[E_n] = [\psi_m]+[\Phi_k]+[\Omega_l] \tag{A.10}$$

REFERENCES

1. L. V. KANTOROVICH and V. I. KEYLOV, *Approximate Methods of Higher Analysis*. Interscience, New York (1958).
2. W. SHOCKLEY, *Electrons and Holes in Semiconductors*. Van Nostrand, New York (1950).
3. P. M. MORSE and H. FESHBACK, *Methods of Theoretical Physics* Vol. II. McGraw-Hill, New York (1953).
4. D. P. KENNEDY, *J. Appl. Phys.* **31**, 1 (1960).
5. *IBM Technical Report* (00.749) (1961).
6. D. P. KENNEDY, *IBM J. Res. Dev.* **5**, 1 (1961).
7. D. P. KENNEDY, *J. Appl. Phys.* **31**, 6 (1960).
8. H. S. CARSLAW and J. C. JAEGER, *Conduction of Heat in Solids*. Oxford University Press, New York (1947).
9. G. N. WATSON, *A Treatise on the Theory of Bessel Functions*. Cambridge University Press, London (1945).

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 226–232. Printed in Great Britain

THE SELECTIVE MELTING OF GERMANIUM BY THERMAL RADIATION

T. C. TAYLOR and C. J. BARDSLEY

Raytheon Company, Semiconductor Division, Newton, Massachusetts

(*Received* 25 *May* 1961; *in revised form* 13 *June* 1961)

Abstract—The selective partial melting of Ge by means of radiant heating is described, wherein a furnace providing a one-dimensional thermal gradient is used in conjunction with means for altering the radiant heat flux to local areas of a Ge sample surface. An analytical model is developed to describe, qualitatively, the mechanics of the heat-transfer processes involved. Experimental results are presented to show that selective local melts in Ge can be produced with the aid of locally deposited carbon particle films on the Ge surface which alter the apparent thermal emissivity, (absorptivity) or with the aid of small, separate bodies of appropriate physical properties resting on the Ge sample surface. Methods are described for introducing doping impurities to the melts, in order to form junction structures.

Résumé—On décrit la fonte partielle sélective du Ge produite par chaleur radiante où un fourneau produisant une rampe thermique est utilisé en conjunction à des moyens pour changer le flux de chaleur aux surfaces locales d'un échantillon de Ge. Un modèle analytique est developpé pour décrire qualitativement le mécanisme des procédés de transfert de chaleur impliqués. Les résultats expérimentaux sont presentés pour démontrer que les fontes locales sélectives dans le Ge peuvent être produites à l'aide de pellicules de carbone déposées localement sur la surface du Ge, qui change l'émissivité thermique apparente, ou à l'aide de petits corps separés ayant des propriétes physiques appropriées reposant sur la surface de l'échantillon de Ge. Des méthodes sont décrites pour introduire des impuretés de dope aux fontes pour former des structures de jonction.

Zusammenfassung—Das selektive teilweise Schmelzen von Ge durch Wärmebestrahlung wird beschrieben. Verwendet wird ein Heizofen, der einen eindimensionalen Wärmegradienten erzeugt, zusammen mit einer Vorrichtung, die den Strahlungsfluss auf lokale Gebiete der Oberfläche der Ge-Probe verändern kann. Die auftretenden Wärmeübertragungsprozesse werden durch ein analytisches Modell beschrieben. Die experimentellen Ergebnisse zeigen, dass selektives lokales Schmelzen durch das Aufbringen von Kohlefilmen auf gewisse Stellen der Ge-Oberfläche erzielt werden kann. Dies ändert das scheinbare Emissions- und Absorptionsvermögen. Statt dessen können auch kleine Einzelkörper mit geeigneten physikalischen Eigenschaften auf die Ge-Oberfläche gebracht werden. Es folgt eine Beschreibung von Methoden zur Einführung von Dotierungsmaterial in die Schmelzen zum Zweck der Erzeugung von Übergängen.

I. INTRODUCTION

A VARIETY of junction-forming processes used for making Ge and Si device structures have involved the selective partial melting of small bodies of semiconductor. It is the purpose of this paper to present a new technique for selective partial melting, which may be used to form junction structures, and which may be practised with a simple thermal-radiation furnace. A brief review of the related technology is given, to furnish a background for the present developments in selective melting techniques.

Most of the earlier work on the selective partial melting of semiconductor bodies has been based on the use of a one-dimensional thermal gradient, whereby the material nearest one surface of a chip or slice, or one end of a bar is melted, while the remainder is solid. The junction structures formed by such means are characterized by the controlled depth of penetration of the molten zone, while the other dimensions of the zone are limited only by the boundaries of the semiconductor sample. Processes which are typical have been described by LEHOVEC and BELMONT[1],

LEHOVEC and LEVITAS[2], BILLIG and GASSON[3], PANKOVE[4], and GOODMAN[5]. The changes of conductivity type necessary to produce useful junction structures from these partial melts have been obtained variously by the introduction of impurities to the melt, redistribution of suitably compensating impurities, and diffusion after resolidifying the melt.

Techniques for the production of selective melts having geometries more complicated than one-dimensional have also been described. BILLIG and GASSON[3] have produced a molten annulus on the surface of a cylindrical rod by radio-frequency heating. LEHOVEC[6] has described the generation of controlled areas of melt on larger surfaces by three means: (1) the "engagement" or touching of the semiconductor surface with a heated object, (2) the use of a stream or jet of heated inert gas directed against the area to be melted, and (3) the use of a focused beam of electrons or other electrically charged particles, or a concentrated light beam. Finally, PEARSON and TREUTING[7] produced small local surface melts on silicon which, however, are believed to be due to the presence of minute amounts of alloying impurities which locally depress the melting point.

The method of selective melting and junction formation to be described in the following sections differs from those cited above, and may be characterized briefly as follows.

(1) A thermal gradient furnace is used, such that semiconductor specimens may be placed in a region between a high-temperature radiation heat source and a lower-temperature body, with the specimen resting on the latter. This arrangement is used to produce a thermal gradient in the semiconductor which is primarily one-dimensional.

(2) The surface of the semiconductor nearest the heat source is modified locally in some manner which alters the radiant heat flux to the modified area, causing it to heat to temperatures which differ from those in the surrounding or contiguous areas. By this means, a two-dimensional thermal gradient is imposed in a plane normal to the axis of the original one-dimensional thermal gradient, thereby achieving a three-dimensional gradient. The three-dimensional gradient is sufficient to localize the semiconductor melt both in depth of penetration and in all dimensions of the plane normal to the penetration depth.

II. NOTATION

C	a radiation exchange coefficient, for unit surface area, dimensions of (q/T^4)
R	a thermal resistance, for unit surface area, dimensions of (T/q)
T	absolute temperature
q	a heat flux, dimensions of (energy)/(unit area) (unit time)
θ	a difference of temperatures
1, 2, 3, 4	subscripts used singly and in combination to identify the geometric locations, or paths which with their symbols are associated

III. ANALYTICAL MODEL

It has been proposed that localized melting can be produced on preselected areas of a semiconductor surface by using radiant heating in conjunction with treatments of the semiconductor surface which locally alter the radiant heat flux. As a guide for experimental work, it was desired to construct an analytical model for the heat-transfer processes involved, in order to learn the relative importance of the variables in the process. A complete and rigorous treatment of these processes, which include radiation with some convection and conduction through a gas, and three-dimensional heat conduction in the semiconductor, could become rather tedious. The present purpose is served sufficiently by presenting a qualitative feasibility argument. Accordingly, the heat-transfer processes are treated as one-dimensional, wherein the primary thermal gradient is through the semiconductor piece. The treatment is therefore valid for selective melting only in the limiting case, wherein the transverse thermal gradients result in very small perturbations of the one-dimensional heat flow. The result of such a treatment can, nevertheless, be pertinent to the selective melting of materials which have well-defined melting points, since very small temperature differences suffice to discriminate between the liquid and solid states in such materials.

The development of the analytical model follows, for the geometrical situation illustrated in Fig. 1. For simplicity, the derivation is based

on a steady-state thermal system. Nevertheless, its conclusions should be qualitatively applicable to transient situations.

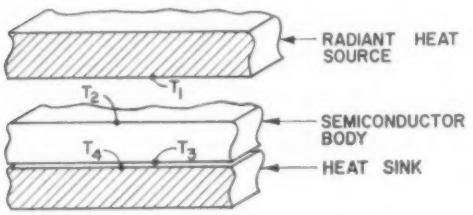


FIG. 1. Schematic representation of the geometry assumed for the analytical model. The radiant-heat source, semiconductor body, and heat sink are visualized as having large, parallel plane faces.

Assumptions

(1) All temperatures of the model system are invariant with time.

(2) The heat source is at a uniform temperature T_1, where T_1 is the highest temperature of the system.

(3) The surface of the heat source is perfectly black, having a radiation emissivity of 1.

(4) The heat sink is at a uniform temperature of T_4, where T_4 is the lowest temperature of the system.

(5) A body of semiconductor is placed between the heat source and the sink, and receives heat from the source through a surface at the uniform temperature T_2, and disposes of heat to the sink from a surface at the uniform temperature T_3. No heat generation occurs within the body.

(6) The principal means of heat transfer from the source to the surface at T_2 is thermal radiation. Other heat transfer processes in the system may be by any means representable by the form $q = (\theta/R)$.

Development

(1) Let the net heat exchange between the heat source and the semiconductor body be represented by, [8]

$$q_{12} = C_{12}(T_1^4 - T_2^4) \tag{1}$$

where the radiation exchange coefficient C_{12} is proportional to the emissivity of the semiconductor surface for the conditions assumed.

(2) Let the heat fluxes through the semiconductor body, and from it to the heat sink be represented by

$$q_{23} = \frac{T_2 - T_3}{R_{23}} \tag{2}$$

and

$$q_{34} = \frac{T_3 - T_4}{R_{34}} \tag{3}$$

(3) By the first law of thermodynamics,

$$q_{12} = q_{23} = q_{34} \tag{4}$$

(4) By appropriate combination of equations (1)–(4),

$$(T_2 - T_4) = C_{12}(T_1^4 - T_2^4)(R_{23} + R_{34}) \tag{5}$$

(5) By differentiation,

$$\frac{\partial T_2}{\partial C_{12}} = \frac{(T_1^4 - T_2^4)(R_{23} + R_{34})}{1 + 4C_{12}(R_{23} + R_{34})T_2^3} \tag{6}$$

From examination of equation (6), the following conclusions are obtained.

(1) For any case where $T_1 > T_2$, $(\partial T_2/\partial C_{12})$ is positive. Hence, if a small area of the semiconductor surface has an emissivity different from the remainder of the surface so that the radiation exchange coefficient differs by ΔC_{12}, the equilibrium temperature of that small area differs from that of the remainder area, where ΔT_2 is of the same sign as ΔC_{12}.

(2) The influence of a given change ΔC_{12} on T_2 is greatest at low emissivity levels.

(3) The influence of a given change ΔC_{12} on T_2 is greatest for large thermal gradients, between the heat source and the semiconductor, since it increases with increasing T_1, and increases with decreasing T_2.

(4) The influence of a given change ΔC_{12} on T_2 is greater for a larger sum of the thermal resistances $(R_{23} + R_{34})$.

It will be noted that the last two effects, stated in conclusions (3) and (4) are in conflict, since a large thermal resistance militates against a large thermal gradient (if T_2 is fixed at the melting point of the semiconductor, large values of thermal resistance coupled with a fixed value of T_4 require that T_1 be small). However, it appears from the form of equation (6) that the variables T_1 and T_2

are far more important in defining the value of $(\partial T_2/\partial C_{12})$ than are the thermal resistances.

The foregoing has been based on the tacit assumption that it is possible to alter the thermal emissivity of a semiconductor surface by some means. In experimental fact, it is more convenient to apply a film of some other material to the surface. In the case of very thin, adherent films, doing this may be considered equivalent to an alteration of the surface emissivity.

The manner of deriving equation (6) is rigorous only for cases where the dimensions of the altered-emissivity area are very small compared to those of the semiconductor as a whole, including its thickness, and for cases where ΔT_2 is small. Otherwise, the assumptions of constant values of the thermal resistances R_{23} and R_{34}, and of a constant temperature T_3 are not justified. Nevertheless, it is found experimentally that selective melting can be done successfully without adherence to the condition that the altered-emissivity area be small.

Since the thermal conductivity of semiconductors such as Ge is relatively large, it is expected that the temperature differences which can be achieved by the mechanism suggested by equation (6) are rather small. The elemental semiconductors have a well-defined melting point, however, so that it is possible to distinguish between the liquid and solid phases if very small temperature differences are obtained. In summary, therefore, it should be possible to produce selective melting by either of two means, namely:

(1) by locally treating the surface to *increase* the emissivity, thereby causing melting to occur only in the *treated* area in the presence of a strong radiant heating field,

(2) by locally treating the surface to *decrease* the emissivity, thereby causing melting to occur only in the untreated area in the presence of a strong radiant heating field.

Consideration of the detailed heat-transfer processes involved in the use of surface films to alter the radiation emissivity of a semiconductor surface leads to another technique for selective melting. It is proposed that instead of a film, the surface treatment be visualized as a separate body (or particle) resting on that surface of the semiconductor which faces the heat source. This body receives heat from the source, and conducts it from its face nearest the source to that nearest the semiconductor. The heat then transfers from the body to the semiconductor by passing through the thermal resistance at their mutual interface. Consideration of this technique was prompted by two facts, namely: (1) certain of the emissivity "film" treatments consist of placing a large number of small particles on the semiconductor surface, and (2) interesting selective melting results are obtained from the use of single homogeneous bodies in lieu of film treatments.

By analyzing the heat transfer processes involved, it can be shown that selective melting may occur under a small body, called a heat receptor, or simply "receptor", provided the body has a suitable combination of the following characteristics:

(1) a large product of surface area and emissivity

(2) a large bulk thermal conductance

(3) a low thermal resistance (good geometrical fit) to the transfer of heat between the body and the underlying semiconductor.

In principle, the opposite of these conditions should produce a "shading" action which prevents melting under a body for conditions which otherwise produce melting.

Section IV presents some experimental results which have been achieved from experiments based on this analysis.

IV. EXPERIMENTAL

Selective melts in a variety of patterns have been produced in Ge by employing a technique based on the emissivity film model. Local areas of a Ge surface were coated with thin carbon-particle films, on the theory that this treatment would be tantamount to an alteration of the surface emissivity. The films may be applied in a slurry form, as by using a colloidal graphite, suspended in a volatile liquid vehicle. The slurry may be applied in preselected patterns by using a masking and painting technique, or with the aid of a pen, or similar ink-handling device. After application, the pattern is dried, and the masking, if any, is removed. Another effective technique is the use of a graphite (common) pencil "lead" to draw the desired pattern on a lapped, or otherwise fine-grained, roughened Ge surface. After such treatment, the Ge is fired in a thermal gradient furnace in an argon atmosphere, under conditions which produce melting under the graphite films, by

using a resistance-heated graphite heat source at a temperature in the range of 1680–1950°C.

Typical results obtained from the pencil-marking technique are shown for four patterns in Fig. 2. In addition to the melting under the pencil lines, it may be noted that some melting (indicated by high shine) also occurs along the upper edges and at the corners of the dice. This fact is due to the enhanced ratio of surface to volume of the dice in these regions. The heat transmitted by radiation to the die is received through the die edges as well as the top surface, while the heat loss to the sink, being mainly by means of gaseous conduction through a thin film of argon between the die and the sink, occurs mostly through the base area of the die. Therefore, one might expect the upper corners and edges of the die to assume higher equilibrium temperatures than the flat upper surface of the die. This edge and corner melting effect was first observed as a problem in preparing surface weld–regrowth junctions.[9] Consideration of the thermal mechanics involved in this problem led to the present investigation of selective melting.

Experiments designed to study the processing of selective melts such as those shown in Fig. 2 provide two results which are of interest in relation to the analytical model (Section III). In the first experiment it was desired to explore the hypothesis that the primary means of heat transfer between the heat source and the samples is by thermal radiation. To do this, selective melt samples were fired in a furnace at several settings of the distance between the lower surface of the heat source and the upper surface of the sample. It was found that well-defined selective melts were obtained at distances of about 0·25 in., while general surface melting nearly obliterated the selective effect of the pencil marks at distances of about 0·05 in. It is concluded from this that at the smaller gap distance, the heat transfer by gaseous conduction and/or convection becomes appreciable. In view of this result, the larger spacing between the heat source and the specimens was used in all other work.

In the second experiment, comparative samples were fired in two different furnaces. The furnaces differed in that one employed water cooling of the heat sink, such that selective melting occurred on the samples when the heat source was maintained at about 1915°C, while the other, having a sink cooled only by convection and radiation, produced selective melting for a heat-source temperature of 1680°C. It was qualitatively observed that the water-cooled furnace yielded a substantially better-defined selective melt, with less tendency for the melt to grow outside of the areas treated with penciled films. This observation is in accord with conclusion (3) [following equation (6)]. Applying equation (6) to this experiment, it is found that the only difference for the two furnaces is represented in the factor $(T_1^4 - T_2^4)$, which indicates that $(\partial T_2/\partial C_{12})$ is about 70 per cent greater for the furnace equipped with a water-cooled heat sink.

Although the analytical models are developed on the assumption of thermal steady state, the samples described may be produced in a very short time. Selective melts such as those shown in Fig. 2 are typically produced by heating from room temperature to surface-melting conditions in times of the order of 20–30 sec. After the melting is produced, the samples may be cooled immediately, or sustained in the selectively melted condition for times of the order of minutes. The stability of the molten regions results from small thermal gradients in the plane of the Ge sample surface, which result in turn from details of the furnace design, the use of radiant heating, and the high thermal diffusivity of Ge. The furnace used for most of the work reported is equipped with a graphite strip as the heat source, having a width of 0·5 in., a thickness of 0·1 in., and a length of 2·5 in. between the electrode jaws. Samples of Ge are fired beneath the center of the strip, where a small hole in the strip permits observation of the samples with a pyrometer telescope. The furnace is powered by a 5 kVA auto-transformer, step-down transformer combination.

As indicated earlier, it is recognized that the application of a carbon-particle film is not, in reality, an alteration of the surface emissivity of Ge. Therefore, experiments were conducted to discover evidence of selective melting which was controlled by means of separate bodies, acting as "receptors" or "shades". A typical result of this search is found in Fig. 3, which shows selective melts on Ge dice which were produced by using smaller silicon dice as heat receptors. The molten Ge near its melting point does not

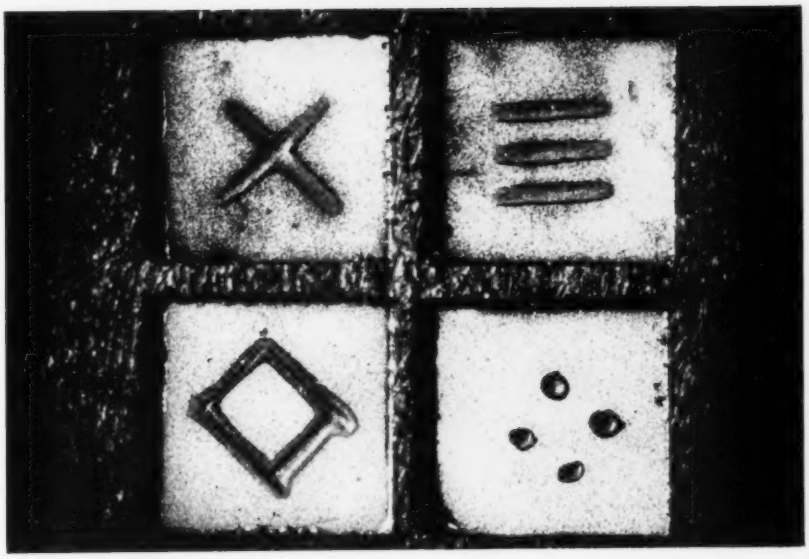


FIG. 2. Selective melt patterns on Ge dice. The patterns were applied to a lapped Ge surface by drawing with pencil, followed by firing under a radiant-heat source. The darker lines which edge the pencil marks are shiny, and are due to growth of the melt beyond the edges of the pencil marks. The Ge dice measure $0{\cdot}125 \times 0{\cdot}125 \times 0{\cdot}011$ in. thick.

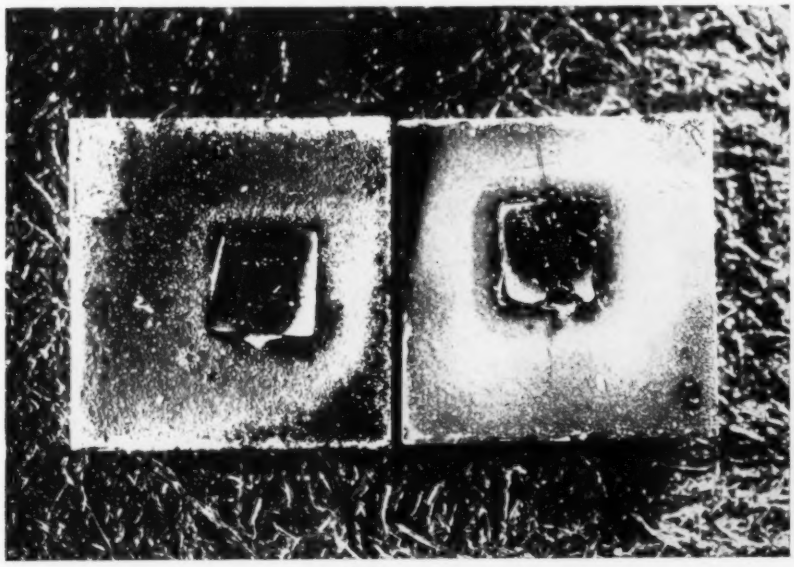


FIG. 3. Selective melt patterns on Ge dice. The patterns were obtained by using Si dice ($0{\cdot}045 \times 0{\cdot}045 \times 0{\cdot}012$ in. thick) as heat receptors, and firing under a radiant heat source. The Ge dice measure $0{\cdot}125 \times 0{\cdot}125 \times 0{\cdot}011$ in. thick.

[*facing page* 230

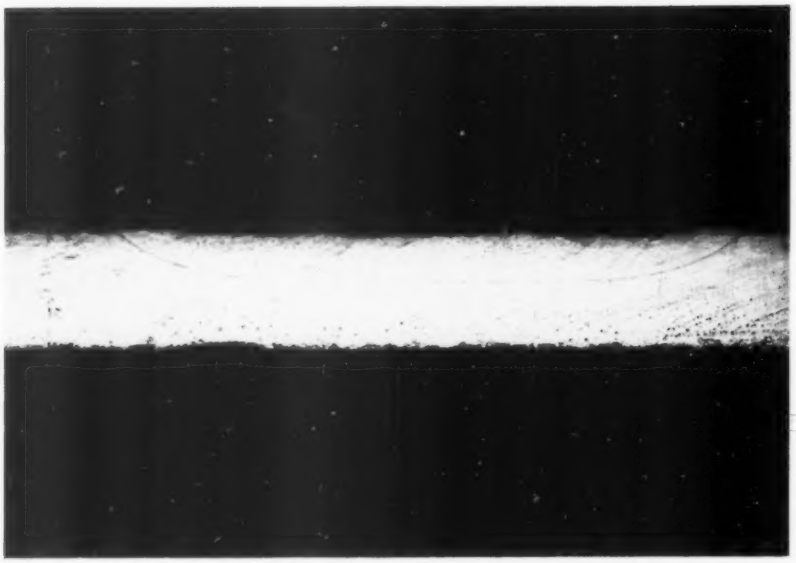


FIG. 4. Photomicrographic section of a receptor-controlled selective melt in Ge, similar to those of Fig. 3. The Ge substrate is *p*-type, In-doped, 0·25–0·35 Ω-cm. resistivity, and 0·012 in. thick. The melt is counter-doped with arsenic, applied before melting in the form of a solution of As_2O_5 dried on the receptor. Note the dish-shaped *p*–*n*-junction line at top. (×105)

wet the Si, so that the receptor is easily removed after firing. With Si receptors, it has been found possible to obtain quite accurate plan-view impressions of the receptor shape in the Ge melt, although the expansion of Ge on freezing usually develops a mound along at least one edge of the melt (note the sharp highlights at the bottom center of the melt regions of Fig. 3). Blocks of graphite may also be used as receptors, although the resulting melt shape is more difficult to control. The experimental conditions which produced the samples of Fig. 3 were very similar to those used for pencil-marked selective melts, except that the Ge surface was etched, since the rough surface texture required to accept a pencil mark is unnecessary in the practice of receptor-controlled melting.

The use of shading to prevent the melting of Ge in local areas has also been explored in a preliminary way. It has been found that thick coatings of finely divided alumina, or barium–strontium–calcium carbonates, may be painted on using a water vehicle, and show evidence of shading action. The results are less impressive than those achieved with receptors, however. It is apparently more difficult to find the combination of low radiation emissivity and thermal conductivity required in a shade material than it is to find the opposite combination desired for a receptor. It is expected, however, that this difficulty may be overcome by the use of reflective coatings on bodies of low thermal conductivity.

Semiconductor junction structures can be made by the selective melting techniques described above. One obvious method of obtaining them is to use impurity segregation effects in compensated Ge. In the present study, methods have been developed for the introduction of doping impurities to the melt during the selective melting process. One method involves the inclusion beneath the receptor of smaller quantities of the desired impurity in a specially prepared chemical matrix. When the receptor is heated as described, the matrix is decomposed, the impurity is liberated, and the freed impurity dissolves rapidly in the melted portion of the substrate.

To overcome the volatility of some commonly used impurities, e.g. arsenic, a specially prepared matrix was designed. This matrix consists of the oxide of the impurity dissolved with the appropriate amount of chelating agent or sequesterant. Specific sequestering agents used were ammonium (dinitrilo)-tetra-acetate and tartaric acid. These have no appreciable amounts of alkali ions or other residual metallic ions present. The amount of dopant introduced was controlled by the impurity concentration of the water solutions and by the amount of matrix included beneath the receptor.

Several variations in the technique of application of the matrix were used. The solution was dried on the receptor or on the substrate, and the solution was included in a graphite film painted on the surface.

In all of these variations the mechanism during heating was the same. The chelating agent decomposed to carbon and gaseous products, then the residual carbon reacted with the impurity oxide, reducing it to the impurity element, which is readily dissolved in the molten zone. The sequesterant inhibits the volatility of the impurity oxide until the decomposition takes place. Since the time of heating is short, and since the chelated bond is relatively stable, the deterrent is sufficient to obtain controlled impurity diffusion without the excessive loss of the dopant.

Other methods include the placing of a small amount of a nonvolatile, powdered impurity under a receptor, using the water solution of the chelated impurity oxide in a suspension of colloidal graphite which is applied to the surface and which acts as an absorptive coating, and marking the surface with a pre-doped pencil lead to form an absorptive doping deposit as in Fig. 2.

The various methods used result in sufficient impurity counter-doping to form *p–n* junctions. These junctions penetrate the substrate to varying distances depending on the particular technique used to form them. An example of the result of using a silicon receptor die on *p*-type germanium and the arsenic doping solution to give an *n*-type doped region beneath the die is shown in Fig. 4. That these are true *p–n* junctions was further established by electrical measurements in the laboratory. High ratios of forward to reverse current of the correct polarities were obtained using samples prepared in the same way as that in Fig. 4.

V. CONCLUSION

The experimental results have shown that it is possible to achieve selective melting of Ge by

comparatively simple means. Specifically, a furnace apparatus, equipped with a radiant-heat source, and designed primarily to produce a one-dimensional thermal gradient, may be used in conjunction with surface treatments, or separate bodies placed on the Ge surface, to produce the three-dimensional thermal gradients required for simple forms of selective melting. The fabrication of junction samples and related experiments indicate that the thermal mechanics involved in selective melting by these means are described qualitatively by the analytical model developed for that purpose.

It seems likely, from the results obtained, that selective melting by means of these and similar techniques will find application in specialized semiconductor-device fabrication and microcircuit fabrication. At present, alternative methods of making complicated device or circuit structures, such as photolithographic techniques, are superior from the standpoints of lower product-size limits and better reproducibility, and this may prove to be the case indefinitely. Nevertheless, the simplicity of the apparatus and versatility of selective melting techniques may result in their being of value in developmental studies, or in short-run production work.

Although the present experimental study has been limited almost exclusively to work with germanium, the basic techniques are, in principle, applicable to the selective melting of other materials, including conductors and insulators as well as semiconductors, provided only that such materials possess a well-defined melting point. Each material however, is likely to present unique technological difficulties. A carbon-film technique cannot be used with Si, for example, apparently because a reaction between C and Si forms SiC at temperatures below those required to effect selective melting. Similarly, attempts to apply the heat-receptor technique to the selective melting of Si have encountered the difficulty that most refractory materials are wetted by molten Si, making removal of the receptor impossible. One material which is found to give reasonably good receptor-controlled melts on Si is boron nitride. Its performance in this capacity is noticeably improved by coating the BN surface which faces the radiant heat source with a carbon-particle film.

Acknowledgements—The writers wish to thank Messrs. G. NAPOLEON and R. DELANEY for help with the experimental work, and Mrs. M. PLUMRIDGE for doing the metallographic and photographic work. The furnace equipment was designed by Mr. F. L. YUAN.

REFERENCES

1. K. LEHOVEC and E. BELMONT, *J. Appl. Phys.* **24,** 1482 (1953).
2. K. LEHOVEC and A. LEVITAS, *J. Appl. Phys.* **28,** 106 (1957).
3. E. BILLIG and D. B. GASSON, *J. Appl. Phys.* **28,** 1242 (1957).
4. J. I. PANKOVE, *Transistors I.* RCA Laboratories, 82 (1956).
5. C. H. L. GOODMAN, *Solid-State Electron.* **1,** 188 (1960).
6. K. LEHOVEC, *U.S. Pat.* 2855334 (1958).
7. G. L. PEARSON and R. G. TREUTING, *Acta Cryst.* **11,** 397 (1958).
8. E. SCHMIDT, *Thermodynamics.* p. 473. Clarendon Press, Oxford (1949).
9. T. C. TAYLOR and F. L. YUAN, *Solid-State Electron.* **3,** 70 (1961).

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 233–238. Printed in Great Britain

AN INVESTIGATION OF $Li_xNi_{1-x}O$ AS A MIXED-VALENCE THERMOELECTRIC MATERIAL

J. C. DANKO, G. R. KILP and H. M. FERRARI

Thermoelectric–Thermionic Development, Atomic Power Department,
Westinghouse Electric Corporation, Pittsburgh 30, Pa.

(*Received* 8 *February* 1961)

Abstract—The mixed valence material, $Li_xNi_{1-x}O$, has been investigated as a potential thermoelectric material. Measurements of the Seebeck coefficient, $\alpha(\mu$ V/°C), electrical resistivity, $\rho(\Omega$-cm), and thermal conductivity, k (W/cm°C) have been made as a function of temperature and lithium concentration. The thermoelectric figure of merit, $Z(\alpha^2/\rho k)$, reaches a value of approximately $1{\cdot}4\times10^{-4}$ at 1100°C for the composition $Li_{0.04}Ni_{0.96}O$.

Résumé—On a examiné la matériau à valence mixte $Li_xNi_{1-x}O$ comme élément thermoélectrique éventuel. Des mesures du coefficient Seebeck, $\alpha(\mu$ V/°C), de résistivité électrique, $\rho(\Omega$-cm) et de conductivité thermale, k(W/cm °C) ont été prises en fonction de la température et de la concentration de lithium. Le facteur de mérite thermoélectroque $Z(\alpha^2/\rho k)$ atteint environ la valeur de $1{,}4\times10^{-4}$ à 1100 °C pour la composition $Li_{0,04}Ni_{0,96}O$.

Zusammenfassung—Die Verbindung $Li_xNi_{1-x}O$, die gemischte Valenzen enthält, wurde auf ihre Eignung als thermoelektrisches Material geprüft. Der Seebeck-Koeffizient $\alpha(\mu$ V/°C), der spezifische Widerstand $\rho(\Omega$-cm) und die Wärmeleitfähigkeit k (W/cm °C) wurden als Funktionen der Temperatur und des Lithiumgehalts gemessen. Die thermoelektrische Leistungsziffer $Z(\alpha^2/\rho k)$ beträgt bei 1100 °C etwa $1{,}4\times10^{-4}$ bei der Zusammensetzung $Li_{0,04}Ni_{0,96}O$.

INTRODUCTION

DURING the past few years, a number of mixed-valence compounds have been proposed as thermoelectric materials.[1] The thermal and electrical characteristics of these materials can be controlled by varying the chemical composition, and it appears possible to produce materials capable of efficiently converting thermal power directly to electrical power at operating temperatures higher than those of the prsent semiconductor materials.

The transition metals (e.g. Mn, Fe, Ce, Ni and Cu) are characterized by their multivalency, i.e. M^{+2} and M^{+3}, where M represents any of the foregoing elements. If one of these metals exists in a compound such as an oxide, and a univalent element of approximately the same ionic radius is added to the oxide, the valence state of the transition metal ion nearest the substituted ion is disturbed. This perturbation gives rise to a "mixed-valence" compound.

Consider the example of $Ni^{+2}O^{-2}$. If monovalent metal ions M^+ are substituted for a small fraction x of the nickel ions, a mixed-valence compound of the composition $M^+Ni^{+2}_{1-2x}Ni^{+3}_xO^{-2}$ is formed. The smaller charge of the M^+ ions balances the excess charge of the Ni^{+3} ions without the simultaneous introduction of vacancies in the cation lattice. The defect center may be described as an impurity cation of lower relative charge on a cation site plus a positive hole bound on a neighboring host cation. The valence change so produced leads to the production of carriers in the material. The above material, $M_xNi_{1-x}O$, is designated as a p-type mixed-valence or controlled-valence compound because of the positive hole produced by the M-ion substitution. On the other hand, if a univalent ion like M^{+3} is used as the substitution material, the charge is carried by a migrating electron, and an n-type mixed-valence compound is obtained.

Efficient thermoelectric materials are characterized by a high Seebeck coefficient (α,μ V/°C), low electrical resistivity (ρ Ω-cm), and low thermal conductivity (k, W/cm°C). These parameters vary with the compound, and, for a given compound, α and ρ are a function of the concentration of the substituted atoms. The preceding variables are related by IOFFE's figure of merit, defined as

$$Z = \alpha^2/\rho k$$

The above equation indicates the effectiveness of any material in a thermoelectric generator.[2]

For this equation, it is apparent that high values of α and low values of ρ and k are desirable.

PREPARATION OF $Li_xNi_{1-x}O$

The mixed-valence characteristics of nickel oxide can be obtained readily by doping with a univalent material. The most appropriate doping material from ionic radii considerations is Li^+, which can go into solid solution in the nickel oxide lattice up to concentrations of 50 atomic per cent or more.[3]

Lithium-doped nickel oxide pellets were usually prepared by the following reaction:

$$\frac{x}{2}\,Li_2O_2 + (1-x)NiO \rightarrow Li_xNi_{1-x}O$$

The reaction was performed in air at 900°C for 16 hr. After preparation by this reaction, the pellets were pulverized, re-pressed and resintered at 1200°C for 4 hr.

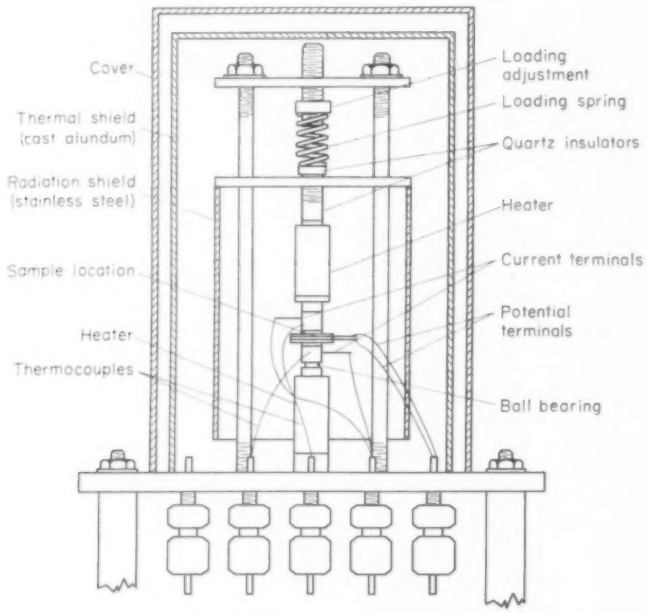


FIG. 1. Schematic diagram of high-temperature apparatus for measuring Seebeck coefficient, electrical resistivity and thermal conductivity.

Pellets containing various amounts of Li^+ displayed an excellent physical appearance as prepared in this manner.

THERMOELECTRIC MEASURING APPARATUS

Fig. 1 shows the apparatus currently used to measure electrical and thermal conductivities and the Seebeck coefficient at elevated temperatures. Since NiO reduces to Ni metal in an inert or reducing environment, the apparatus was designed

so as to operate to 800°C in air. The sample measured consisted of a cylinder of $\frac{1}{2}$ in. diameter and $\frac{1}{4}$ in. to $\frac{1}{2}$ in. length. The cylinders which hold the current terminals and make contact with the sample are made of Monel, as is the core for the upper heater. The three supporting legs, the radiation shield and the disk to which the shield is attached, are stainless steel. The core for the lower heater is copper; the base plate and the upper disk assembly, brass. Electrical connections are made through the base plate by means of Stupakoff seals and Conax connectors. The cover for the apparatus is $\frac{3}{8}$ in. aluminum.

The lower contact cylinder rests on a ball bearing, which corrects for non-parallelism of the sample faces and insures uniform contact. The faces of the samples are metallic-coated, either by evaporation or electroplating. The thermocouples are set into wells in the contact cylinders and held in position by set screws.

To measure the Seebeck coefficient, the e.m.f. at the current terminals is read. To measure resistance, the e.m.f. across the potential terminals for a known current is read. The potential terminals consist of small (0–80) screws held by metal rings cemented to a ceramic cylinder which surrounds the sample. All potentials are measured on a Leeds and Northrup type K–3 potentiometer.

Fot thermal-conductivity measurements, a quartz cylinder is placed in series with the sample, a Monel disk (with thermocouple) is placed between the two, and thermal conductivities are determined by comparing the temperature drops across the two cylinders. The potential terminal assembly is not required, and a Monel cylinder (with holes for the thermocouple leads) connects the two contact cylinders to serve as an inner radiation shield.

For measurements which were conducted on one sample at 1100°C, a modified instrument utilizing platinum-wound heaters and platinum and Monel parts was employed.

THERMOELECTRIC-PROPERTY MEASUREMENTS

Electrical resistivity

Results of electrical resistivity measurements as a function of lithium concentration and the average temperature across the sample are presented in Fig. 2. These data indicate that the electrical resistivity decreases rapidly with an increase in temperature (over most of the temperature range) following the relationship

$$\rho = Ae^{-Q/RT}$$

where A = constant, R = gas constant, Q = activation energy (cal/mole) and T = absolute temperature (°K).

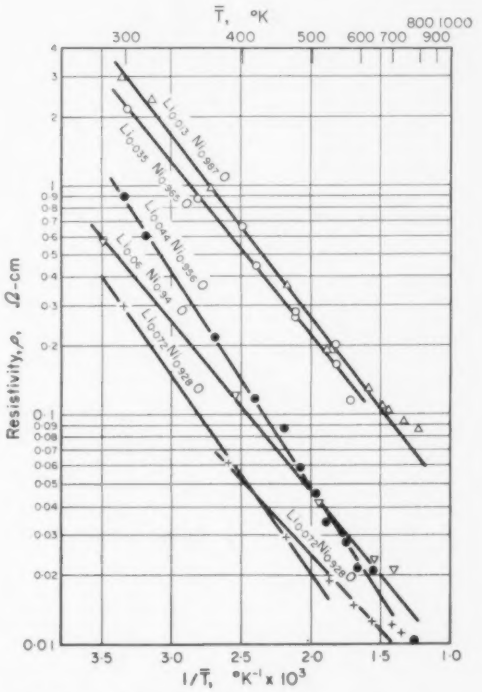


FIG. 2. Electrical resistivity of $Li_xNi_{1-x}O$ as a function of temperature.

Table 1 lists the value of Q for each composition as obtained from the slope of the curve ln ρ vs. $1/T$.

Table 1. *Activation energy* (Q) *for several compositions of* $Li_xNi_{1-x}O$

Composition	Q (cal/mole)
$Li_{0.013}Ni_{0.987}O$	3613
$Li_{0.035}Ni_{0.965}O$	3449
$Li_{0.044}Ni_{0.956}O$	4278
$Li_{0.06}Ni_{0.94}O$	3369
$Li_{0.072}Ni_{0.928}O$	3928

Since the average temperature used in the $\ln\rho$ vs. $1/T$ plots is an approximation determined mathematically from the measured hot and cold temperatures across the sample, plots of $\ln\rho$ vs. the hot temperature T_H and the cold temperature T_C were made as shown in Fig. 3 for $Li_{0.044}Ni_{0.956}O$.

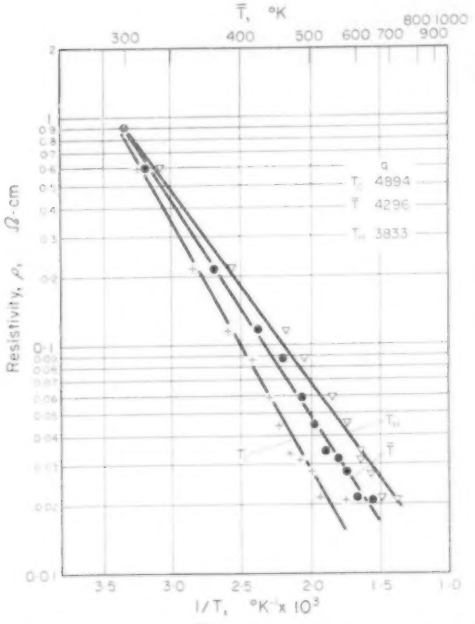


FIG. 3. Ln ρ vs. T_H, $\bar{T}$ and T_C for $Li_{0.044}Ni_{0.956}O$.

From the slopes of these curves, values of Q are obtained which represent the maximum acceptable deviation of the slope of the average temperature ($\bar{T}$) curve used to calculate Q. The results obtained indicated that the Q values should be accurate within 10 per cent.

The qualitative values of Q, which decrease from 4·3 kcal for $Li_{0.044}Ni_{0.956}O$ to 3·4 kcal for $Li_{0.06}Ni_{0.94}O$, are in general agreement with values published by VERWEY[4] and JOHNSON and HEIKES[5]. This activation energy was interpreted by VERWEY as the energy required to free a positive hole from the neighborhood of the lithium ion. Once free, hole motion takes place, without activation, from site to site. MILLER and HEIKES[6], on the other hand, propose a somewhat different model for the conduction process: they visualize the measured activation energy as mainly for carrier mobility and not carrier formation. This model of conduction is supported by recent experimental results.

From Fig. 2, the electrical resistivity decreases as the lithium concentration is increased. This occurs because an increase in lithium ions produces more Ni^{+3} ions and thus a proportionately larger number of positive carriers. The deviation of resistivity from linearity at high temperatures suggests a possible change in the mechanism of conduction in this range.

Seebeck coefficient

From Fig. 4 it is seen that the Seebeck coefficient decreased as the concentration of Li^+ ions (concentration of carriers) increased. Since the electrical properties are also a function of the carrier concentration, the optimization of Z, the figure of merit, will be accomplished by optimizing the Seebeck coefficient and electrical resistivity through adjustments in the doping level.

A rigorous solid-state treatment of the transport properties of mixed-valence materials has been made by HEIKES[7]; the reader is referred to this article for the derivation of the Seebeck coefficient.

Thermal conductivity

For thermoelectric materials, the total thermal conductivity K_T is the sum of k_l, the heat carried by the lattice vibrations; k_e, that carried by electrons (or holes); and k_r, that by radiation, as follows

$$K_T = k_l + k_e + k_r$$

Heat conducted by the carriers is related to the electrical conductivity by the Wiedeman–Franz relationship, and is an important factor in the conventional semiconductor thermoelectric materials. However, for the narrow-band mixed-valence LiNiO material, HEIKES[7] has proposed that the electronic contribution to the thermal conductivity is zero. This effect has not been verified as yet by experimental data.

Fig. 5 is a plot of the thermal conductivity of $Li_xNi_{1-x}O$ as a function of temperature. The data were obtained from a sample in which $x = 0{\cdot}035$, but experience has shown that the thermal conductivity is essentially independent of the doping concentration at the levels investigated in this study.

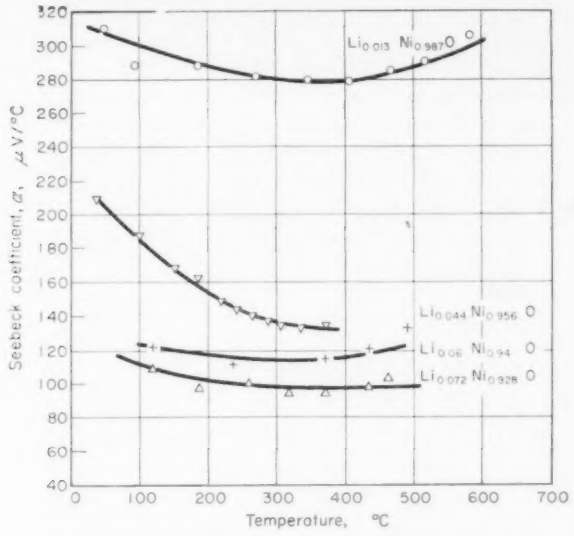


FIG. 4. Seebeck coefficient for $Li_xNi_{1-x}O$ as a function of temperature.

Figure of merit

Since the lithium concentration affects both the electrical resistivity and Seebeck coefficient, there is, therefore, an optimum doping level which results in maximizing $\alpha^2/\rho k$. This is confirmed in Fig. 6, a plot of Z vs. lithium concentrations for isotherms of 300, 400 and 500°C. The figure of merit for these materials is low at low temperatures, but increases with temperature. With $Li_xNi_{1-x}O$, the optimum doping occurs at $x = 0{\cdot}04$; this sample, which exhibited the largest Z value, had a Seebeck coefficient of approximately 200 μ V/°C, which is in agreement with theory.[1]

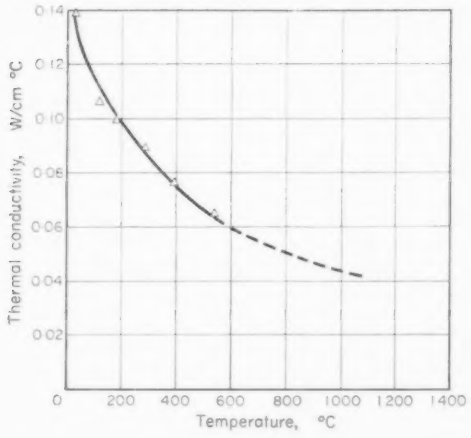


FIG. 5. Thermal conductivity for $Li_{0.035}Ni_{0.965}O$ as a function of temperature.

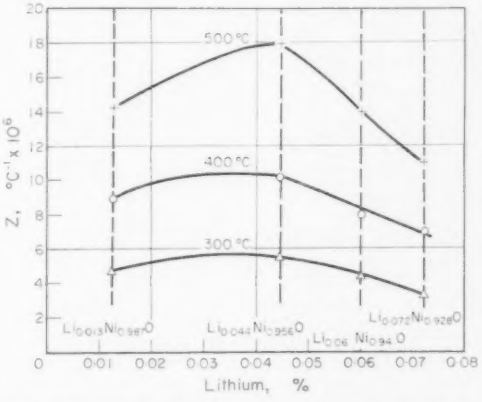


FIG. 6. Figure of merit as a function of composition and temperature for $Li_xNi_{1-x}O$.

$Na_xNi_{1-x}O$ MATERIALS

Pellets of NiO base material doped with Na^+ were also prepared to determine if, in addition to the properties obtained with Li^+ doping, it was possible to obtain the added advantage of being able to attain higher temperatures before sublimation of the doping additive occurred. Higher temperatures are desirable from two viewpoints: (1) the figure of merit increases with increasing temperature, and (2) the Carnot efficiency increases with increasing temperature difference. High-density pellets of the composition $Na_{0.05}Ni_{0.95}O$ were prepared. These pellets had room temperature thermoelectric properties similar to $Li_{0.05}Ni_{0.95}O$. However, the results of a test conducted with a thermobalance indicated that Na_2O began to volatilize at approximately 1000°C; therefore, this material is not quite as stable in air as $Li_xNi_{1-x}O$, which can be heated to 1100°C before appreciable volatilization occurs.

CONCLUSIONS

$Li_xNi_{1-x}O$, possessing a figure of merit Z of $1{\cdot}4 \times 10^{-4}$ at 1100°C, shows some potential as a thermoelectric material. It must, however, be used in a controlled oxidizing environment in the range 800–1100°C. The optimum value of x (Li^+ ions) lies at approximately 0·04.

Acknowledgements—The authors wish to express their appreciation to Dr. G. L. THOMAS for taking the measurements; to the Westinghouse Research Laboratories for measurements of one sample at very high temperatures; and to the U.S. Atomic Energy Commission for sponsoring this study.

REFERENCES

1. C. ZENER, *Thermoelectricity* **9,** 30 (1958).
2. A. F. IOFFE, *Semiconductor Thermoelements and Thermoelectric Cooling*. Infosearch Limited, London (1957).
3. J. B. GOODENOUGH, D. G. WICKHAM and W. J. CROFT, *J. Phys. Chem. Solids* **5,** 107 (1958).
4. E. J. W. VERWEY, *Philips Res. Rep.* **5,** 173 (1950).
5. W. D. JOHNSON and R. R. HEIKES, *J. Amer. Chem. Soc.* **78,** 3255 (1956).
6. R. C. MILLER and R. R. HEIKES, *J. Chem. Phys.* **28,** 398 (1958).
7. R. R. HEIKES, *Science and Engineering of Thermoelectricity*. Interscience, New York (1960).

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 239–260. Printed in Great Britain

STUDY OF THE FAST TRANSIENT BEHAVIOUR OF PELTIER JUNCTIONS

K. LANDECKER and A. W. FINDLAY

Department of Physics, University of New England
Armidale, N.S.W., Australia

(*Received* 1 *March* 1961; *in revised form* 28 *June* 1961)

Abstract—A method has been developed for measuring the temperature of a thermojunction after the passage of fast transient current pulses (30 msec to several seconds duration).

Using the model of a thermocouple whose branches are of infinite extent, it is shown theoretically that for rectangular current pulses the temperature of the cold junction is a function only of a parameter $\zeta = I\sqrt{\tau}$ (I = pulse current, τ = pulse duration). The lowest temperature occurs for one particular value $\zeta_{\min}$ of this parameter, and this temperature cannot be reduced further by any particular choice of the individual values of I and τ. In connexion with these rectangular pulses, various forms of the temperature dependence of the Seebeck coefficient are examined. For the case of logarithmic dependence, a method is developed which makes possible the determination of the temperature distribution along the couple from a measurement of the Seebeck potential as a function of time and allows the influence of the Thomson effect to be estimated.

It is further shown that, when a second rectangular current pulse is superimposed on to the first current pulse, the temperature drop due to the combined pulses may be at most twice that caused by a single pulse.

Finally it is proved, with the aid of a variational method, that the lowest transient temperature drops of the junction may be obtained with pulses for which the current is a continuous function of the time, of the form

$$Z(\Theta)_{ex} = \text{const.}\ (t-\Theta)^{-1/2},\ t > \Theta$$

where Θ is the time and t the time of observation.

If the pulse terminates at time t_1 and if $t \to t_1$ then $Z(\Theta)_{ex} \to \infty$. Provided the current were allowed to rise indefinitely, the transient temperature would approach absolute zero if the process were not counteracted by changes in the Seebeck coefficient and the electrical and thermal conductivity which all have been assumed to be constant in this calculation.

The above results are also true for "real" thermocouples provided the length of the branches is $l \gg (2\kappa\tau)^{1/2}$ where κ is the (average) diffusivity of the thermoelectric substances.

These calculations were essentially confirmed by experiments. With the rising current pulses, the accuracy of measurement is limited by the time resolution of the measuring circuit and by the fact that the temperature dependence of the Seebeck coefficient in this region is not known accurately. However, it is almost certain that we have observed transient temperatures well below 100°K.

Résumé—Une méthode a été développée pour mesurer la température d'une jonction thermique après le passage d'impulsions transitoires de courant très rapides (la durée variant de 30 msec à plusieurs secondes.)

En utilisant un modèle de thermocouple dont les branches s'étendent indéfiniment, il a été prouvé théoriquement que pour des impulsions rectangulaires de courant, la température de la jonction froide est en seule fonction du paramètre $\zeta = I\sqrt{\tau}$ (I = courant d'impulsion et τ = durée d'impulsion). La température minimum se produit à une valeur particulière $\zeta_{\min}$ de ce paramètre, et elle ne peut plus être réduite par un choix des valeurs individuelles de I et τ. En rapport avec ces impulsions rectangulaires les diverses formes que prend le coefficient de Seebeck en fonction de la température sont examinées. Dans le cas d'une fonction logarithmique, une technique pour évaluer la distribution de température le long du couple en mesurant la tension Seebeck en fonction de la durée a été développée. Elle permet aussi d'estimer l'influence de l'effet Thompson.

Il est ensuite démontré que lorsqu'une seconde impulsion est superimposée sur la première la chute de température des deux impulsions ne peut être plus que deux fois plus grande que celle causée par une impulsion.

Finalement on prouve à l'aide d'une méthode variationelle que des chutes transitoires minima de température à la jonction peuvent être obtenues avec des impulsions où le courant est en fonction continue de la durée telle que

$$Z(\Theta)_{ex} = \text{const.}\ (t-\Theta)^{-1/2},\ t > \Theta,$$

ou Θ indique la durée et t la durée d'observation.

Si l'impulsion se termine à un temps t_1 et si $t \to t_1$, alors $Z(\Theta)_{ex} \to \infty$. En assumant une montée de courant illimitée, la tempérautre transitoire approcherait le zero absolue si le procédé n'aurait pas été contrecarré par certains changements du coefficient Seebeck et la conductibilité thermique et électrique qu'on a assumées constantes dans ce calcul.

Les resultats ci-haut sont aussi applicables pour des thermocouples réels pourvu que la longueur des branches soit $l \gg (2\kappa\tau)^{1/2}$, où κ est le taux de diffusion moyen des matériaux thermoelectriques.

Ces calculs ont été confirmés par des expériences. Dans le cas des montées d'impulsions de courant, la précision est limitée par la résolution de temps des circuits de mesure et par le fait que les changements de coefficient Seebeck en fonction de la température ne sont pas encore connus. Néanmoins il est presque certain que des températures transitoires bien au-dessous de 100 °K ont été observées.

Zusammenfassung—Die Temperatur einer Thermolötstelle nach Durchgang von schnellen Stromimpulsen (30 msek bis mehrere sekunden Dauer) wurde gemessen.

Unter Benutzung des Modells eines Themroelements mit unendlich langem Leiterpaar wird theoretisch nachgewiesen, dass für Rechteckimpulse die Temperatur der kalten Lötstelle nur von einem Parameter $\zeta = I\sqrt{\tau}$ abhängt. (I = Impulsstrom, τ = Impulsdauer). Das Temperaturminimum tritt bei einem bestimmten ζ–Wert auf, und es lässt sich auf keine Weise durch Veränderung der I- and τ-Werte weiter reduzieren. Im Zusammenhang mit den Rechteckimpulsen werden die verschiedenen Formen der Temperaturabhängigkeit des Seebeck-Koeffizienten untersucht. Im Fall einer logarithmischen Temperaturabhängigkeit ergibt sich eine Methode, durch die sich die Temperaturverteilung längs des Elements aus einer Messung des Seebeck-Potentials als Funktion der Zeit ergibt, und die ausserdem eine Abschätzung des Einflusses des Thomson-Effekts ermöglicht.

Es wird weiterhin gezeigt, dass die Überlagerung eines zweiten Rechteckimpulses über den ersten Impuls eine Abnahme der Temperatur hervorruft, die höchstens das Doppelte der von einem Impuls erzeugten Temperaturabnahme ist.

Schliesslich wird mit Hilfe einer Variationsmethode nachgewiesen, dass die grösste Temperaturerniedrigung der Lötstelle mit Impulsen erzielt wird, bei denen der Strom eine kontinuierliche Funktion der Zeit ist:

$$Z(\Theta)_{ex} = \text{const.}\ (t-\Theta)^{-1/2},\ t > \Theta,$$

wo Θ = Zeit, t = Beobachtungszeit.

Endet der Impuls zu einer Zeit t_1, so wird bei $t \to t_1$ $Z(\Theta)_{ex} \to \infty$. Bei unbeschränktem Stromanstieg würde die Durchgangstemperatur nach der absoluten Null streben; dies wird jedoch durch Veränderungen des Seebeck-Koeffizienten und der elektrischen und Wärmeleitfähigkeit verhindert, die alle bei der Berechnung als konstant angesehen wurden.

Die obigen Ergebnisse sind für "wirkliche" Thermoelemente richtig, wenn die Länge der Leiter $l \gg (2\kappa\tau)^{1/2}$ ist, wo κ der durchschnittliche Diffusionskoeffizient der thermoelektrischen Substanzen ist.

Die Berechnungen liessen sich im wesentlichen experimentell bestätigen. Mit steigenden Stromimpulsen wird die Messgenauigkeit durch die Zeitauflösung des Messkreises begrenzt und durch die Tatsache, dass die Temperaturabhängigkeit des Seebeck-Koeffizienten in diesem Bereich nicht genau bekannt ist. Es ist jedoch ziemlich sicher, dass wir Durchgangstemperaturen weit unter 100 °K beobachtet haben.

1. INTRODUCTION

WHEREAS most of the classical work on the Peltier effect was concerned with thermojunctions through which a stationary or quasi-stationary current flows, some anomalous behaviour of these junctions was occasionally observed[1] immediately after the switching on of the current, but nothing beyond a qualitative explanation was given. BOERDIJK[2] points out that the transient behaviour of thermocouples depends on their shape and dimensions, in contradistinction to couples with stationary currents. The only detailed studies of the transient behaviour of Peltier junctions that have come to our notice are those of STIL'BANS

and FEDOROVICH[3] and more recently of PARROTT[4] and of GRAY[5]. The first two authors theoretically and experimentally investigated the temperature distribution after the switching on of the current through a Peltier junction. They also showed that by superimposing on the stationary current a current pulse in the shape of a step function, a considerable excess cooling could be obtained during a short time interval. PARROTT[4] made a similar analysis of the stationary as well as of the transient behaviour of refrigerating couples taking into account the Thomson effect and the variation with temperature of the electrical conductivity and of the surface heat transfer. The study of GRAY[5] is mainly concerned with the control characteristics of thermoelectric heat pumps and generators, e.g. with the effects on their behaviour of changes in current, loading and ambient temperature.

In the present work we are interested in time intervals very much shorter than envisaged by these authors. The times involved here are usually of the order of a small fraction of a second. The following calculations were in fact carried out in connexion with the use of an experimental method which allows the recording of very fast variations of the Seebeck potential of a junction and which is described in Section 3 below.

We first summarize the assumptions that have been made in these calculations.

Because the accuracy of measurement is not very high the electrical conductivity has been assumed to be independent of temperature. The thermal conductivity must also be taken as constant over the temperature range considered in order to make the calculations manageable. Moreover, the model of the thermojunction referred to in the following Section is based on the additional assumption that the electrical conductivity and the thermal diffusivity in the n- and p- branches are equal. This will always be at least approximately true for practical thermoelectric substances.

Where, in the following, the Peltier coefficient is not taken to be a constant, the Seebeck coefficient (thermoelectric power) has been assumed to be independent of temperature except in one case considered below where logarithmic variation has been taken into account.

Though the results obtained are also applicable to Peltier heating it must be emphasized that our main interest is in the cooling effect. To simplify the exposition we have nevertheless referred throughout the text to "Peltier heat" and it should be noted that we visualize this as a negative quantity.

2. DERIVATION OF THE WAVEFORMS AND MAGNITUDES OF FAST TRANSIENTS OF TEMPERATURE IN THERMOCOUPLES DUE TO CURRENT PULSES UNDER VARIOUS CONDITIONS

Notation

Following is a list of the more important symbols used in this Section:

v	temperature due to Peltier and/or Thomson effect (°C)
V	temperature due to combined Peltier (Thomson) and Joule effect (°C)
T	absolute temperature (°K)
T_0	absolute temperature of the hot junction (°K)
q	rate of flow of heat (cal/sec-cm^2)
s	Laplace variable
$\mathscr{L}\{\xi\}$	Laplace transform of ξ
$\mathscr{L}^{-1}\{\eta\}$	inverse Laplace transform of η
v^*	Laplace transform of v
q^*	Laplace transform of q
x	co-ordinate of length
t, θ	co-ordinates of time
τ	duration of rectangular current pulse
$\bar{\tau}$	oscillographically determined pulse duration
I	pulse current
$\bar{I}$	oscillographically determined constant pulse current
$\zeta = I\sqrt{\tau}$	a parameter
a	cross-sectional area of thermocouple
l	length of thermocouple branches
K	thermal conductivity (cal/sec-cm-deg)
σ	electrical conductivity ((Ω-cm)$^{-1}$
ρ	density (g/cm^3)
c	specific heat (cal/g-deg)
$\kappa = K/\rho c$	diffusivity (cm^2/sec)
S	Seebeck coefficient (thermoelectric power)†
P	Peltier coefficient
γ	Thomson coefficient
ϕ	thermoelectric (Seebeck) potential
M	strength of delta function $\delta(t)$ of quantity of heat (cal.)
N	height of step function of heat flux when P = const. (cal/sec)
Q	pulse height of heat flux when $P = ST$, S = const. (cal/sec.)
R	pulse height of heat flux when $P = ST$, $S = r \ln T + C_0$ (cal/sec.)
F	staircase function of heat flux (cal/sec.)
Y	staircase function of current (A)

†Terminology recommended by the Joint Task Group on Thermoelectric Devices of the A.I.E.E. and I.R.E.

$G = \frac{1}{2}P/4{\cdot}2$ conversion factor ($4{\cdot}2 \simeq$ Joules equivalent) (cal/A-sec.)
$Z(\Theta)$ an arbitrary current wave form (A)
$Z(\Theta)_{ex}$ an extremal of $Z(\Theta)$ (A)

We use the model of the thermocouple shown in Fig. 1.

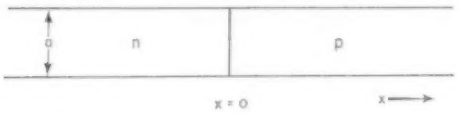


FIG. 1. Schematic diagram of the model of a thermocouple (a = cross-sectional area).

Two semi-infinite rods, n and p, of thermoelectric substances and of uniform cross-sectional area a extend along the x-axis and are joined at the origin without any intervening layer.[3–5] At the time $t = 0$, Peltier heat is produced at the origin by switching on currents of various magnitudes lasting up to various times $t = \tau$. The free surfaces of the rods are assumed to be perfectly insulated against loss or gain of heat.

The problem of calculating the temperature distribution along the branches may be greatly simplified if in first approximation the Peltier coefficient P is assumed to be a constant. Then the effects of the Peltier heat and Joule heat may be considered separately. The justification of this procedure is most easily expressed in terms of the delta-function solution of Subsection A. Since the production of Joule heat is uniform throughout the volume of the rods and is always concomitant with the production of Peltier heat, the separation of the two heats is certainly justified for a delta-function type of current pulse. Then it is also justified for a current pulse of arbitrary form, since the latter may always be thought of as a superposition of delta-function "responses". When the Peltier coefficient is a function of temperature, as in Subsections C and D, the separation of the Joule heats is not possible, that is, a heat production term must be included in the equation of heat conduction. However, when the Peltier coefficient is a linear function of temperature, as in Subsection C, the first method may still be applied as an approximation, as is shown in that section. The heat-production term is introduced into the basic equations in subsection D.

To calculate the effect of the Peltier heat we consider the half of the thermocouple extending from the junction ($x = 0$) in the direction of the positive x-axis towards infinity. The temperature $v(x,t)$ and the heat current $q(x,t)$ along this semi-infinite thermal "transmission line" are completely determined by the heat-conduction equations

$$\frac{\partial v}{\partial t} = \kappa \frac{\partial^2 v}{\partial x^2} \tag{1a}$$

$$\frac{\partial q}{\partial t} = \kappa \frac{\partial^2 q}{\partial x^2} \tag{1b}$$

or the equivalent pair

$$\frac{\partial v}{\partial t} = -\frac{\kappa}{aK}\frac{\partial q}{\partial x} \tag{2a}$$

$$q = -aK\frac{\partial v}{\partial x} \tag{2b}$$

To solve these equations we use the method of the Laplace transformation. The Laplace-transformed set corresponding to (1) is

$$\frac{d^2v^*}{dx^2} - \frac{s}{\kappa}v^* = -\frac{1}{\kappa}v(x,0) \tag{3a}$$

$$\frac{d^2q^*}{dx^2} - \frac{s}{\kappa}q^* = \frac{aK}{\kappa}\frac{\partial}{\partial x}v(x,0) \tag{3b}$$

where s is the Laplace variable and v^*, q^* are the Laplace transforms of $v(x,t)$ and $q(x,t)$ respectively.

The Laplace-transformed set of the pair (2) is

$$v^* = -\frac{\kappa}{aKs}\frac{dq^*}{dx} + \frac{1}{s}v(x,0) \tag{4a}$$

$$q^* = -aK\frac{dv^*}{dx} \tag{4b}$$

Since the whole thermocouple is at the temperature of the hot junction until time $t = 0$, it is natural to adopt this temperature as the zero of the temperature scale. Then the right-hand side of equations (3ab) and the last term in (4a) are always zero. If there are reasons to adopt any other temperature zero, the additional complication in obtaining solutions is admittedly only trivial, but the homogeneity of the transformed equations effected by this simple artifice greatly simplifies our notation.

In all of the following calculations the temperature of the junction is not given as a function of time. More specifically, the heat input due to the Peltier and Thomson effects is itself a function of temperature. Although the form of this function is given by the Kelvin relations, its exact nature can only be determined by reference to a specific atomic model of the substance or by experiment in any particular case. In the following we examine various forms of the temperature dependence of the Peltier coefficient. Since in semiconductors such as bismuth telluride, which we have used, the sign of the Thomson coefficient is such that it enhances the Peltier cooling over the temperature range of interest, we consider initially the Peltier effect alone. The Thomson effect will be considered later.

A. *The "delta-function" solution*

We begin with the simple case where the current pulse applied to the thermocouple is of very short duration so that the Peltier heat generated may be considered to be in the nature of a delta function.

A quantity of heat of magnitude

$$M = \tfrac{1}{2} P\bar{I}\bar{\tau}/4{\cdot}2 \qquad \text{(cal)} \tag{5}$$

may then be thought of as being released instantaneously at the junction. Here P is the Peltier coefficient, $\bar{I}$ is the constant current flowing during a short time interval $\bar{\tau}$; both $\bar{I}$ and $\bar{\tau}$ are experimental quantities which may be read from an oscillogram of the current pulse. The factor 4·2 is Joule's equivalent and the factor $\frac{1}{2}$ indicates that only one half of the Peltier heat flows into the branch, the other half flowing into the left-hand branch of the couple. Then the input function is

$$q(0,t) = M \cdot \delta(t)$$

where $\delta(t)$ is the delta function and M is the strength of the delta function. If, for simplicity, we initially assume P to be a constant, the solution for $v(x,t)$ and $q(x,t)$ may be written down immediately.

The solution of equation (3b) is

$$q^*(x,s) = \mathscr{L}\{q(0,t)\}\, e^{-x\sqrt{(s/\kappa)}} \tag{6}$$

which in the present case becomes

$$q^*(x,s) = \mathscr{L}\{M \cdot \delta(t)\} e^{-x\sqrt{(s/\kappa)}} = Me^{-x\sqrt{(s/\kappa)}} \tag{6a}$$

since the Laplace transform of the delta function is unity. Therefore

$$q(x,t) = \mathscr{L}^{-1}\{Me^{-x\sqrt{(s/\kappa)}}\} = M\mathscr{L}^{-1}\{e^{-x\sqrt{(s/\kappa)}}\}$$
$$= \tfrac{1}{2}M(\pi\kappa)^{-1/2}xt^{-3/2}e^{-x2/4\kappa t} \tag{7}$$

using the inverse transform

$$\mathscr{L}^{-1}\{e^{-x\sqrt{s}}\} = \frac{x}{2\sqrt{(\pi t^3)}}\, e^{-x^2/4t} \text{ for } \begin{matrix} x > 0 \\ t > 0 \end{matrix} \tag{8}$$

H.C.P. No. 82†

and the transformation formula

$$\mathscr{L}^{-1}\{F(s/\kappa)\} = \kappa f(t \cdot \kappa) \tag{9}$$

Then from (2b) follows by integration the temperature distribution

$$v(x,t) = M(aK)^{-1}(\kappa/\pi)^{1/2}t^{-1/2}e^{-x^2/4\kappa t} \tag{10}$$

This result may also be obtained from (4a) using the transform

$$\mathscr{L}^{-1}\left\{\frac{e^{-x\sqrt{s}}}{\sqrt{s}}\right\} = \frac{1}{\sqrt{\pi t}}\, e^{-x^2/4t} \text{ for } t > 0 \tag{11}$$

H.C.P. No. 84

and formula (9).

The class of functions of the form (7) and (10) is well known in the theory of heat conduction. The quantities q and v tend to zero for all values of x as $t \to 0$ except at the origin where they become infinite. They nevertheless are true representations of heat flow and temperature at all x and all t with the exception of a small neighbourhood near the origin, determined by the conditions of a particular experiment. In our case the limitations are quite clear: nothing can be said about q and v within the time interval $\bar{\tau}$ which is of the order of a small fraction of a second and which is determined unambiguously by an oscillographic record. That M in equation (5) is the correct expression of the instantaneously released quantity of heat to be used in (10) may be confirmed by integrating $v(x,t)$ over the entire branch, obtaining

$$\int_0^\infty v(x,t)\rho ca\,dx = (\rho ca)M(aK)^{-1}(\kappa/\pi)^{1/2}t^{-1/2}$$
$$\times \int_0^\infty e^{-x^2/4\kappa t}dx = M \tag{12}$$

† The sources of the transforms used in this note are denoted by the abbreviations H.C.P., E. and C.J. which refer to Refs. 6, 7 and 8 respectively.

For a given point x, the temperature is a maximum at a time

$$t_{\max} = x^2/2\kappa \tag{13}$$

and the value of v at this time is

$$v(x,t_{\max}) = M(2/e\pi)^{1/2}(aK)^{-1}\kappa x^{-1} \tag{14}$$

We further note that for a given time t' the root-mean-square distance of the heat wave from the origin is

$$x_{\text{r.m.s.}} = (2\kappa t')^{1/2} \tag{15}$$

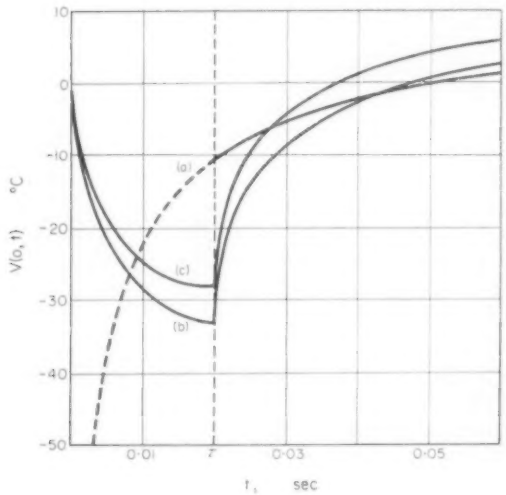


FIG. 2. Calculated temperature $V(0,t)$ at the junction.
(a) Delta-function current pulse,
(b) current of finite duration τ, P = const. and
(c) current of finite duration τ, $P = ST$, S = const.
(The curves are drawn for couple No. 7 and correspond to a pulse current of 77 A and a pulse duration of 20 msec.).

Equation (15) allows us to estimate the extent to which the following results will be applicable to conditions in real thermocouples. For example, with a time of observation of 1 sec and a value of $\kappa \simeq 5\times10^{-3}$ cm^2 sec^{-1} (appropriate to bismuth telluride) the results will be significant for real couples if their length is at least several times the r.m.s. distance of $\sim$ 1 mm.

When the Joule heat produced in the couple is taken into account, remembering that M must be a negative quantity for Peltier cooling, it follows that the temperature distribution becomes

$$V(x,t) = \bar{I}^2\bar{\tau}/4{\cdot}2a^2\sigma\rho c - M(aK)^{-1}(\kappa/\pi)^{1/2}t^{-1/2}e^{-x^2/4\kappa t} \tag{16}$$

Although the function $v(x,t)$ is unbounded at the origin ($x = 0$) [see curve a, Fig. 2], $V(0,t)$ as a function of current has a minimum for any given (finite) time. The lowest possible minimum that may be observed will be at the end of the experi-

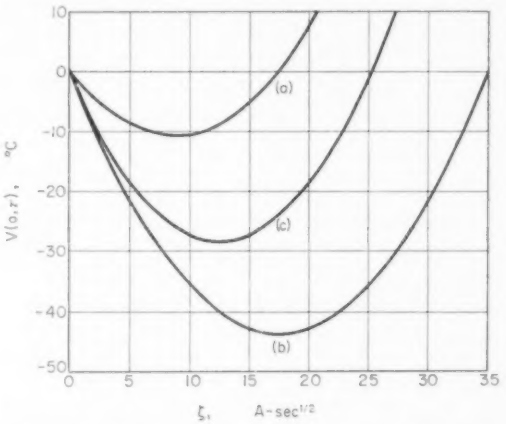


FIG. 3. Calculated temperature $V(0,\tau)$ at the junction as a function of $\zeta = I\sqrt{\tau}$ (see text). Curves (a), (b), (c) respectively correspond to the same conditions as those for Fig. 2.

mental time $\bar{\tau}$ because this is the earliest time at which the Seebeck potential can be measured. We therefore have, from (16) and (5),

$$V(0,\bar{\tau}) = \bar{I}^2\bar{\tau}/4{\cdot}2a^2\sigma\rho c - \tfrac{1}{2}(aK)^{-1}(\kappa/\pi)^{1/2}P\bar{I}\bar{\tau}^{1/2}/4{\cdot}2 \tag{17}$$

We note that, in this expression, current and time occur in the combination $\zeta = \bar{I}\sqrt{\bar{\tau}}$ only. The lowest observable temperature will then occur for a value

$$\zeta_{\min} = \tfrac{1}{4}\,\sigma a(\kappa\pi)^{-1/2}P \tag{18}$$

and have a magnitude

$$V(o,\bar{\tau})_{\min} = -\frac{1}{16}\,\frac{\pi^{-1}}{4{\cdot}2}\,\frac{\sigma}{K}\,P^2 \tag{19}$$

These relations are illustrated in Fig. 3, curve a.

If we now drop the assumption that the magnitude of the Peltier coefficient P is constant it turns out that no advantage is gained by restricting ourselves to a delta-function type of pulse. The analysis of a current pulse of finite duration is actually simpler. However, in order to enable us to estimate the effect of the transition from a delta-function type of pulse to a pulse of finite width, we retain in the first step the assumption that P is constant.

B. *Pulse of finite duration* τ, (P = *const.*)

In our experimental arrangement described in Section 3, the thermocouple was most frequently connected to a constant-voltage supply by means of a switch so that the current through the couple rose instantaneously to a value determined solely by the circuit resistance and remained constant until the switch was opened. This is of course the situation usually met in practice. If we then initially assume that the Peltier coefficient P is constant, i.e. independent of temperature as in Subsection A, the heat flux into the junction is also constant while the current flows. The pulse of heat flux may therefore be thought of as being the superposition of two step functions of height

$$N = \tfrac{1}{2}\,PI/4{\cdot}2 \quad \text{(cals/sec)} \tag{20}$$

where $P = S_0T_0 = \text{const.}$

If the unit step function at $t = 0$ is denoted by $U(t)$, equation (6) becomes

$$q^*(x,s) = \mathscr{L}\{N[U(t)-U(t-\tau)]\}e^{-x\sqrt{(s/\kappa)}} = \frac{N}{s}(1-e^{-\tau s})e^{-x\sqrt{(s/\kappa)}} \tag{21}$$

since

$$\mathscr{L}\{U(t)\} = \tfrac{1}{s} \text{ and } \mathscr{L}\{U(t-\tau)\} = \frac{e^{-\tau s}}{s}$$

Using the transformation formula

$$\mathscr{L}^{-1}\{F(s)e^{-\tau s}\} = f(t-\tau)\,U(t-\tau) \tag{22}$$

(21) transforms to

$$q(x,t) = N[\operatorname{erfc}(x/2\sqrt{\kappa t})-\operatorname{erfc}(x/2\sqrt{\kappa(t-\tau)}\times U(t-\tau)] \tag{23}$$

C.J. No. 8

Substituting (21) into (4a) and using (9) there results

$$v^*(x,s) = \frac{N}{aK\kappa}\left[\frac{e^{-x\sqrt{(s/\kappa)}}}{(s/\kappa)^{3/2}} - \frac{e^{-x\sqrt{(s/\kappa)-\tau s}}}{(s/\kappa)^{3/2}}\right] \tag{24}$$

and with the inverse transform

$$\mathscr{L}^{-1}\left\{\frac{e^{-\alpha^{1/2}s^{1/2}}}{s^{3/2}}\right\} = 2\pi^{-1/2}t^{1/2}e^{-\alpha/4t} - \alpha^{1/2}\operatorname{erfc}(\tfrac{1}{2}\alpha^{1/2}t^{-1/2}) \tag{25}$$

E. No. 7 p. 246

$x = \sqrt{\alpha}$ and formula (9), we finally obtain

$$v(x,t) = \frac{N}{aK}[2(\kappa/\pi)^{1/2}t^{1/2}e^{-x^2/4\kappa t} - x\operatorname{erfc}[x/2(\kappa t)^{1/2}] - \{2(\kappa/\pi)^{1/2}(t-\tau)^{1/2}e^{-x^2/4\kappa(t-\tau)} - x\operatorname{erfc}(x/2[\kappa(t-\tau)]^{1/2}\}U(t-\tau)] \tag{26}$$

This result can also be obtained by substituting (23) into (2b) and integrating from 0 to x.

Since N must be negative for Peltier cooling it follows that the combined effect of Peltier cooling and Joule heating for the complete pulse gives rise to a temperature

$$V(x,t) = [btU(t)-b(t-\tau)U(t-\tau)]-v(x,t) \tag{27}$$

where $b = I^2/4{\cdot}2a^2\sigma\rho c.$

If the temperature V is looked upon as a function of the pulse duration ($t = \tau$) this becomes, at the origin ($x = 0$) (cf. curve b, Fig. 2),

$$V(0,\tau) = I^2\tau/4{\cdot}2a^2\sigma\rho c-(aK)^{-1}(\kappa/\pi)^{1/2}PI\tau^{1/2}/4{\cdot}2 \tag{28}$$

We note that, in this expression also, current and time occur in the combination $\zeta = I\sqrt{\tau}$ just as we previously found in expression (17). The minimum of $V(0,\tau)$ occurs for the value

$$\zeta_{\min} = \tfrac{1}{2}\sigma a(\kappa\pi)^{-1/2}P \tag{29}$$

and the lowest temperature is therefore

$$V(0,\tau)_{\min} = -\tfrac{1}{4}\frac{\pi^{-1}}{4{\cdot}2}\frac{\sigma}{K}P^2 \tag{30}$$

ζ_{min} and $V(0,\tau)_{min}$ are seen to be respectively twice and four times the corresponding values found previously for the delta-function pulse under similar conditions [cf. equations (18) and (19) and curves a and b of Fig. 3]. Remembering that the temperature of the junction due to a rectangular pulse may be thought of as an integration from 0 to τ over a sequence of delta-function "responses" this is precisely the result we would expect.†

Using the numerical values for bismuth telluride from the list quoted below it is found that the lowest temperature has the value

$$V(0,\tau)_{min} = -\tfrac{1}{4}\frac{(10{\cdot}5\times10^{-2})^2\times10^3}{\pi\times4{\cdot}2(4{\cdot}8\times10^{-3})} = -43{\cdot}5^0\text{C}$$

It is of interest to compare this figure with the maximum temperature drop obtainable with a "real" thermocouple when the cross-sectional area and length are given their optimum values and when optimum current flows through the couple. With the well-known assumption that one half of the Joule heat flows to the cold junction[9, 10] this maximum temperature drop is given by

$$\Delta V = \tfrac{1}{2}zT_c^2$$

where the "figure of merit"

$$z = \tfrac{1}{4}S^2\frac{\sigma}{K}. \qquad (31)$$

Using the same values for the thermoelectric properties as before and assuming a temperature of the cold junction of $T_c = 250\ ^\circ$K, gives

$$\Delta V \simeq 47{\cdot}5\ ^\circ\text{C}.$$

The magnitude of the temperature drop due to impulse cooling is therefore found to be about equal to that for cooling by a stationary current. However, it must be kept in mind that with an infinitely long branch, strictly speaking "stationary" cooling does not occur. Even the smallest current will ultimately cause the temperature of the junction to rise again if it persists for a time exceeding τ in an expression such as

$$\zeta_{min} = \text{const.} = \tfrac{1}{2}\sigma a(\kappa\pi)^{-1/2}P = I\sqrt{\tau}$$

no matter how long the time interval τ may be. We shall refer to this fact again in Subsection E.

C. *Pulse of finite duration* τ, ($P = ST$, $S = const.$)

We now allow for the fact that, according to the first Kelvin relation, the Peltier coefficient must be of the form $P = ST$, but we initially assume S to be a constant. Then the pulse applied at the junction has the value

$$Q = \tfrac{1}{2}(ST_0+Sv)I/4{\cdot}2 \quad (\text{cal/sec}) \qquad (32)$$

i.e. the heat flux into the juntion is now a function of temperature. Putting $I'' = \frac{1}{2}I/4{\cdot}2$ as an abbreviation, the solution of (3a) has to satisfy the boundary condition

$$q(0,t) = -aK\left(\frac{\partial v}{\partial x}\right)_{x=0} = -Q = -I''(ST_0+Sv)$$

or

$$\left(\frac{\partial v}{\partial x}\right)_{x=0} = (I''/aK)(ST_0+Sv) \qquad (33)$$

This is seen to be of the form of the well-known "radiation boundary condition" in the theory of heat conduction, and the solution is as follows: The Laplace transformed equation (33) is

$$\left(\frac{dv^*}{dx}\right)_{x=0} - (I''S/aK)v^* = (I''ST_0)/aKs \qquad (34)$$

and the solution of equation (3a) that satisfies this condition is

$$v^*(x,s) = -\frac{(I''S/aK)T_0e^{-x\sqrt{(s/\kappa)}}}{s[(s/\kappa)^{1/2}+(I''S/aK)]} \qquad (35)$$

The inverse transform of the form

$$\frac{hT_0\,e^{-x\sqrt{(s/\kappa)}}}{s[\sqrt{(s/\kappa)}+h]}$$

corresponding to (35) is well known and the temperature due to the pulse up to a time t is found to be

$$v(x,t) = -T_0\{\text{erfc}[x/2(\kappa t)^{1/2}] - e^{hx+h^2\kappa t}\,\text{erfc}[x/2(\kappa t)^{1/2}+h(\kappa t)^{1/2}]\} \qquad (36)$$

C.J. No. 14

† Note:

$$\int_0^\tau [M(aK)^{-1}(\kappa/\pi)^{1/2}]\,U(\theta)\cdot(t-\theta)^{-1/2}d\theta = 2[M(aK)^{-1}(\kappa/\pi)^{1/2}]U(\theta)\cdot\tau^{1/2} \text{ when } t=\tau.$$

where $h = (I'' S/aK)$, so that the temperature distribution as a consequence of the pulse of length τ (curve c, Fig. 2) generating Peltier heat and in the presence of simultaneous Joule heating becomes approximately

$$V(x,t) \simeq [btU(t)-b(t-\tau)U(t-\tau)]+v(x,t) \quad (37)$$

where $b = I^2/4{\cdot}2a^2\sigma\rho c$. The validity of this approximation will be discussed later in this section and an exact solution will be given.

Considered as a function of pulse duration, (37) gives rise to a temperature at the origin

$$V(0,\tau) \simeq I^2\tau/4{\cdot}2a^2\sigma\rho c$$
$$-T_0\{1-e^{(I''S/aK)^2\kappa t}\operatorname{erfc}[(I''S/aK)(\kappa t)^{1/2}]\} \quad (38)$$

We note that even in this case current and time occur only in the combination $\zeta = I\sqrt{\tau}$ so that (38) may also be written as

$$V(0,\tau) \simeq \zeta^2/4{\cdot}2a^2\sigma\rho c - T_0[1-e^{\lambda^2\zeta^2}\operatorname{erfc}(\lambda\zeta)] \quad (39)$$

where

$$\lambda = S\kappa^{1/2}/8{\cdot}4aK$$

but here $\zeta_{\min}$ and therefore also $V(0,\tau)_{\min}$, cannot be computed easily and are best determined graphically (curve c, Fig. 3.). Use of the same numerical values of σ, K and κ resulted in a magnitude of $V(0,\tau)_{\min} \simeq 28°C$, i.e. somewhat smaller than in the previous case where P was assumed to be independent of temperature. Also $V(0,\tau)_{\min}$ turned out to be independent of the cross-sectional area a as before.

We now examine the effect of the approximation involved in using equation (37). Since the Peltier coefficient is a function of temperature, the input function Q [equation (32)] is affected by the Joule heat as well as by the Peltier heat, and the junction temperature should strictly be derived from the basic equation (45), of Subsection D. Then it is found that (37) must be replaced by

$$\mathscr{L}^{-1}\{V^*(x,s)\}$$
$$= \mathscr{L}^{-1}\{W\kappa/Ks^2+v^*(x,s)(1+W\kappa/T_0Ks)\} \quad (37a)$$

where W is the Joule heat produced per unit time and unit volume and $v^*(x,s)$ is given by (35). The term

$$\mathscr{L}^{-1}\{v^*(x,s)W\kappa/T_0ks\}$$

represents a correction on $V(x,t)$ which makes this temperature exact. This derivation is entirely analogous to the calculation of $V_2(x,t)$ equation (54) and will not be repeated here. However, it is remarkable that, in spite of the comparative complexity of this correction term, the current and the time still appear only in the by now familiar form of the parameter ζ when $x = 0$. We therefore give the junction temperature here in that form:

$$V(0,\tau) = \zeta^2/4{\cdot}2a^2\sigma\rho c - T_0[1-e^{\lambda^2\zeta^2}\operatorname{erfc}(\lambda\zeta)]$$
$$+16{\cdot}8(K/\sigma S^2)\{e^{\lambda^2\zeta^2}\operatorname{erfc}(\lambda\zeta) \quad (39a)$$
$$-\sum_{p=0}^{2}[\lambda\zeta]^p/\Gamma(\tfrac{1}{2}p+1)\}$$

where Γ denotes the gamma-function. When numerical values appropriate to bismuth telluride are inserted into this equation it is found that the correction to $V(0,\tau)_{\min}$ amounts to no more than a few per cent. The minimum junction temperature in fact turns out to be slightly lower than the approximate calculation would indicate.

D. *Pulse of finite duration* τ ($P = ST$, $S = r \ln T + C_0$, *temperature region* 150–300°K)

If S is an arbitrary function of temperature, a solution cannot in general be obtained easily. However, it has recently been shown[11] that the Seebeck coefficient of n- and p-type bismuth telluride closely obeys a law of the form

$$|S_{n,p}| = r \ln T + C_0 \quad (40)$$

between 150°K and room temperature which is the region we are particularly interested in.

We shall now show that this logarithmic dependence of the Seebeck coefficient on temperature allows us to obtain the temperature distribution along the thermocouple by numerical computation and also to include the influence of the Thomson effect on the temperature distribution.

As outlined in connexion with our experimental arrangement in the next Section, the quantity we actually measure is not the temperature at the junction but the Seebeck potential which arises from that temperature. For current pulses of constant strength I and varying durations $\tau_1, \tau_2, \tau_3 \ldots$, we obtain an experimental curve $\phi = f(t)$ by measuring the peak amplitudes of oscillographic records like the one shown in Fig. 7(b). All of the temperature curves of Figs. 9(a, b) through 11(a, b) were derived from such ϕ-curves.

Let a heat pulse of magnitude

$$R = I''S(T)T = I''2(r\ln T + C_0)T \text{ (cal/sec)} \quad (41)$$

be applied at the input to the branch where I'' is the current defined above and r is known from experiments as described in Ref. 11. Then in general

$$dR = I''(SdT + TdS) = I''\left(SdT + T\,\frac{dS}{dT}\,dT \right) \quad (42)$$

and with $d\phi = SdT$, and in the special case where $S = 2(r\ln T + C_0)$, this becomes

$$dR = I''(d\phi + 2rdT) \quad (43)$$

or, on integration,

$$R = I''[(\phi - \phi_0) + 2r(T - T_0)] = I''[\phi + 2rV] \quad (44)$$

since $\phi_0 = 0$ and $(T - T_0) = V$.

In contradistinction to the cases treated previously, the input function (44) contains, in addition to the temperature, the Seebeck potential arising from the two causes of Peltier heating and Joule heating. In order to find the respective contributions of these two heats we must add to equation (1a) a term which takes account of the production of Joule heat in the branches. Equation (1a) then becomes

$$\frac{\partial^2 V}{\partial x^2} - \frac{1}{\kappa}\frac{\partial V}{\partial t} = -\frac{W}{K} \quad (45)$$

Here $W = I^2/4{\cdot}2\,a^2\sigma$ is the Joule heat in calories produced per unit time and per unit volume. The boundary condition appropriate to the (negative) heat pulse described by (41) and (44) is

$$\left.\begin{aligned} \left(\frac{\partial V}{\partial x}\right)_{x=0} &= (I''/aK)\phi + 2r(I''/aK)V \text{ or} \\ \left(\frac{\partial V}{\partial x}\right)_{x=0} - 2nrV &= n\phi \end{aligned}\right\} \quad (46)$$

where we have used the abbreviation $n = I''/aK$. This is seen to be again of the form of the radiation boundary condition except for the presence of the factor ϕ. We now imagine that $\phi = f(t)$ may be approximated by an analytical expression of arbitrary form which has a Laplace transform

$$\phi^* = \mathscr{L}\{\phi(t)\} \quad (47)$$

Then the Laplace-transformed equations (45) and (46) become respectively

$$\frac{d^2V^*}{dx^2} - \frac{s}{\kappa}V^* + \frac{1}{\kappa}V(x,0) = -\frac{W}{Ks} \quad (48)$$

and

$$\left(\frac{dV^*}{dx}\right)_{x=0} - 2nrV^* = n\phi^* \quad (49)$$

In (48) the term $(1/\kappa)V(x,0) = 0$ as a consequence of the particular choice of the temperature zero discussed before, and so the solution of (48) and (49) is

$$\begin{aligned} V^*(x,s) &= -\frac{n\phi^* e^{-x\sqrt{(s/\kappa)}}}{[(s/\kappa)^{1/2} + 2nr]} + \frac{2nrW\kappa e^{-x\sqrt{(s/\kappa)}}}{Ks^2[(s/\kappa)^{1/2} + 2nr]} + \frac{W\kappa}{Ks^2} \\ &= V_1^*(x,s) + V_2^*(x,s) + V_3^*(x,s) \end{aligned} \quad (50)$$

We now discuss in turn the derivation of the inverse Laplace transform of the three terms V_1^*, V_2^* and V_3^* in equation (50).

If in V_1^* we put

$$\chi^* = -\frac{ne^{-x\sqrt{(s/\kappa)}}}{[(s/\kappa)^{1/2} + 2nr]} \quad (51)$$

then, by the convolution theorem,

$$V_1(x,t) = \mathscr{L}^{-1}\{\phi^* \cdot \chi^*\} = \int_0^t \phi(t-\tau)\chi(\tau)d\tau \quad (52)$$

The inverse transform of (51) is known to be

$$\begin{aligned} \chi = &-n\{(\kappa/\pi t)^{1/2}e^{-x^2/4\kappa t} \\ &-2nr\kappa e^{2nrx + (2nr)^2\kappa t}\,\mathrm{erfc}[x/2(\kappa t)^{1/2} + 2nr(\kappa t)^{1/2}]\} \end{aligned} \quad (53)$$

C.J. No. 12

The contribution V_1 to the temperature $V(x,t)$ is therefore obtained by convolving the experimental curve $\phi = f(t)$ with the function χ given by (53) for every x and every t. The graphical determination of V_1 for one particular pair of values of x and t is illustrated in Fig. 4. However, when $V(x,t)$ is required for an extended region of the variables and for various functions ϕ (t) (corresponding to different currents), the evaluation of the convolu-

tion integral (52) is hardly practical without some computational aids. A relatively rapid method for the numerical computation of such integrals has recently been described by JORDAN[12].

The inverse transform of the term V_2^* may be derived from a tabulated form and has the value

$$V_2(x,t) = (W\kappa^2/K(2nr)^2) \times \{e^{2nrx+\kappa t(2nr)^2}\,\mathrm{erfc}[x/(2\sqrt{\kappa t})+2nr\sqrt{\kappa t}] - \sum_{p=0}^{2}[-4nr\sqrt{\kappa t}]^p i^p\,\mathrm{erfc}(x/2\sqrt{\kappa t})\} \quad (54)$$

C.J. No. 16

where

$$i^p\,\mathrm{erfc}\,\eta = \int_{\eta}^{\infty} i^{(p-1)}\mathrm{erfc}\xi d\xi$$

and

$$i^0\,\mathrm{erfc}\,\eta = \mathrm{erfc}\,\eta$$

denote p-fold repeated integrations over the complementary error function. These integrals are tabulated in Ref. 8. It is of interest that V_2 does not involve ϕ but represents the modification of the Peltier coefficient by the Joule heat.

The inverse transform of the third term V_3^* is

$$V_3(x,t) = (W\kappa/K)t = I^2t/4{\cdot}2a^2\sigma\rho c. \quad (55)$$

This term is independent of x and represents as before the temperature due to Joule heat accumulated up to the time t.

We finally show that for short transient currents, and in this special case where S is given by (40), the Thomson effect may be taken into account very simply.

When thermocouples are operated under stationary conditions it is usually assumed, for the purpose of calculating the efficiency, that the Thomson heat may be thought of as contributing one half of its value to each junction.[13,14] It is difficult to calculate the exact fraction of the Thomson heat which, after a given lapse of time, has diffused to the cold junction. However, this fraction must of necessity approach unity for a pulse of vanishing duration. With pulses of short but finite duration the choice of this fraction will depend on whether the interest is in the temperature of the junction or in the temperature distribution along the branches. If the temperature is to be determined when x and t have appreciable values, then the entire Thomson heat must be added to the Peltier heat because it is immaterial whether the heat input is into the junction surface or into the bulk of a small layer containing the junction. Furthermore, our experimental results lead us to believe that, with short transient pulses, even the junction temperature is more correctly accounted for when the above fraction is chosen closer to unity than to the value one half customarily used with stationary cooling. The Thomson heat in bismuth telluride is, however, on the average so small that the problem is here of minor importance (cf. the remarks in Section 3C). With this assumption

$$R = q_P + q_T = I''(ST + \int \gamma dT) \quad (56)$$

where q_P and q_T are the Peltier and Thomson

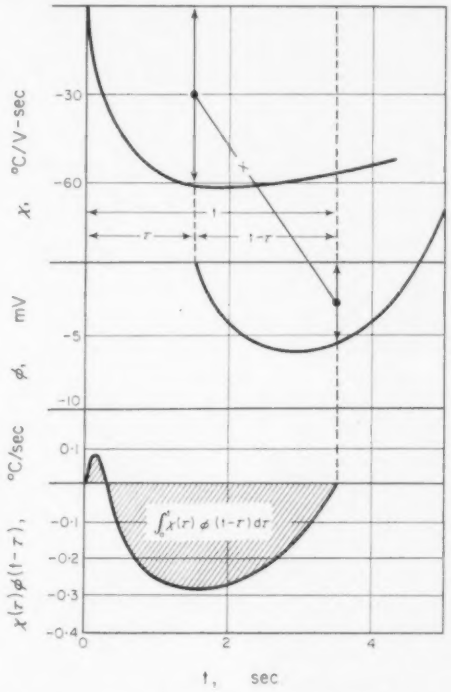


FIG. 4. Diagram illustrating the derivation of the temperature distribution $V_1(x,t)$ by convolution of the Seebeck potential ϕ with the χ-function equation (53) (see text).

heats respectively and the Thomson coefficient γ is given by the second Kelvin relation

$$\gamma = T\frac{dS}{dT} \tag{57}$$

We therefore have

$$dR = I''\left(SdT+2T\frac{dS}{dT}dT\right) \tag{58}$$

With

$$S = 2(r\ln T+C_0)$$

this becomes

$$R = I''(\phi+4r(T-T_0)) = I''(\phi+4rV) \tag{59}$$

The Thomson effect is therefore taken care of by simply using $4r$ instead of $2r$ in all equations following (43).

E. *The superposition of two step-function current pulses and the general "staircase" input function*

We now examine the experiment of STIL'BANS and FEDOROVICH[3] where a step-function-type current pulse is superimposed on to the optimum steady current through a thermocouple after stationary conditions have been established. Under these circumstances the authors observed a transient temperature 30 per cent lower than with optimum stationary cooling. Such enhanced cooling by superposition of step-function currents has also been observed in this and no doubt in other laboratories and these experiments suggest that still lower transient temperatures may be observed if the current through the couple increases in time in the manner of a "staircase function".

Let

$$F = G\,Y(t) \tag{60}$$

be this function where the factor

$$G = \tfrac{1}{2}P/4{\cdot}2 \tag{71}\dagger$$

converts current into calories per second input into the junction, and the Peltier coefficient P is assumed to be constant. Guided by the result obtained with rectangular pulses, namely, that the junction temperature is a function only of the parameter $\zeta = I\sqrt{\tau}$ we define the staircase function $Y(t)$ in (60) as follows:‡

Current "rise"

$$I_\nu = \zeta_s/\sqrt{\theta_\nu} \tag{61}$$

Time "run"

$$\begin{aligned}\theta_\nu &= p^\nu\tau\\ \nu &= 0,1,2\ldots\mu\end{aligned} \tag{62}$$

Here $\zeta_s = \zeta_{\min} = \text{const.}$ is assumed to be determined from formula (29) or by an experiment performed with single rectangular pulses. The subscript s merely indicates that this value of ζ is used in connexion with the staircase function. The justification for treating ζ_s as a constant will be discussed later. The factor p is an arbitrary multiplier to be determined such that the junction temperature is a minimum after μ steps. When the first time run τ is arbitrarily chosen, all of the current rises I_ν are fixed, provided p can be calculated. The case $\mu = 1$ corresponds to the experiment of STIL'BANS and FEDOROVICH. For this case the temperature of the junction may immediately be written down by observing that the problem of finding the Laplace transform of (60) with $\mu = 1$ becomes identical with that of determining the transform of the expression (21) when in the latter N is replaced by G and the second step function $U(t-\tau)$ is given a positive instead of a negative sign. Further current steps $U(t-\tau-p\tau)$, $U(t-\tau-p\tau-p^2\tau)$, etc. may be added indefinitely by an extension of this method. Omitting the intermediate arithmetical steps in this calculation, the junction temperature due to Peltier heat at a time t is found to be

$$\left.\begin{aligned}v(0,t) = \frac{G\zeta_s}{\sqrt{\tau}}\{&2(aK)^{-1}(\kappa/\pi)^{1/2}t^{1/2}\\ &+(1/\sqrt{p})2(aK)^{-1}(\kappa/\pi)^{1/2}(t-\tau)^{1/2}\\ &+(1/\sqrt{p^2})2(aK)^{-1}(\kappa/\pi)^{1/2}(t-\tau-p\tau)^{1/2}\\ &+\ldots\ \}\end{aligned}\right\} \tag{63}$$

and so the temperature after two current steps becomes

$$\left.\begin{aligned}v(0,\tau+p\tau) = \frac{G\zeta_s}{\sqrt{\tau}}&\{2(aK)^{-1}(\kappa/\pi)^{1/2}(\tau+p\tau)^{1/2}\\ &+(1/\sqrt{p})2(aK)^{-1}(\kappa/\pi)^{1/2}(\tau+p\tau-\tau)^{1/2}\}\\ &= (G\zeta_s 2(aK)^{-1}(\kappa/\pi)^{1/2})[1+(1+p)^{1/2}]\end{aligned}\right\} \tag{64}$$

† cf. Section 2F.

‡ For the terminology used with such functions see, for example, Ref. 15.

The temperature due to the Joule heat generated during time $(\tau+p\tau)$ is

$$\left.\begin{aligned} &v_J(0,\tau+p\tau) \\ &= I_0^2\tau/4{\cdot}2a^2\sigma\rho c+(I_0+I_1)^2p\tau/4{\cdot}2a^2\sigma\rho c \\ &= \zeta_s^2/4{\cdot}2a^2\sigma\rho c+\left(\frac{\zeta_s}{\sqrt{\tau}}+\frac{\zeta_s}{\sqrt{p\tau}}\right)^2 p\tau/4{\cdot}2a^2\sigma\rho c \\ &= (\zeta_s^2/4{\cdot}2a^2\sigma\rho c)[1+(1+p^{1/2})^2] \end{aligned}\right\} \tag{65}$$

The junction temperature due to the combined Peltier and Joule heat is therefore

$$\left.\begin{aligned} V(0,\tau+p\tau) &= (\zeta_s^2/4{\cdot}2a^2\sigma\rho c)[1+(1+p^{1/2})^2] \\ &\quad -G\zeta_s 2(aK)^{-1}(\kappa/\pi)^{1/2}[1+(1+p)^{1/2}]. \end{aligned}\right\} \tag{66}$$

However, if after the first current step the thermocouple produces a temperature drop at all, then

$$G\zeta_s 2(aK)^{-1}(\kappa/\pi)^{1/2} > \zeta_s^2/4{\cdot}2a^2\sigma\rho c \tag{67}$$

Also for any value of p

$$(1+p^{1/2})^2 > (1+p)^{1/2} \tag{68}$$

We therefore conclude that the lowest junction temperature after two current steps is obtained as $p \to 0$, and equation (66) then shows that this temperature is *at most* twice as low as the temperature after a single rectangular step. In other words p should be chosen as small as possible, but consistent with the response time of the temperature measuring instrument employed. A quickly responding instrument necessitates a high current rise during the second and all of the following time runs.

It is difficult to compare these results numerically with those of the Russian workers even when we disregard the fact that the present theory strictly applies only to couples with indefinitely long branches. However, the couple shown in their Fig. 1 appears to have approximately the same dimensions as our couple No. 7 and from the data given in the text and Fig. 3 it may be concluded that the factor p was about 1/30. Using formula (66) it then follows that the minimum temperature due to the superposition of the second current step should have been

$$\begin{aligned} V(0,\tau+p\tau) &= v_J[1+(1+\sqrt{1/30})^2] \\ &\quad -v_P[1+\sqrt{(1+1/30)}] \\ &= 48\times 2{\cdot}4-91\times 2{\cdot}015 = 115-183 = -68^\circ \end{aligned}$$

as compared to -52° actually measured.

The calculation of further steps of the staircase function (60) provides no further difficulties and shows that progressively lower temperatures may be produced by consecutive higher current steps of decreasing durations. However, this calculation is otherwise not very instructive. Although the junction reaches a temperature minimum after the first time run τ for a constant value of the parameter ζ it is by no means certain that this is also the most appropriate moment for initiating the second current step. Expressions like (66) should more correctly be minimized also with respect to ζ_s but this is algebraically prohibitive. Furthermore, such staircase functions of current are difficult to produce in practice. For these reasons it is more informative to enquire into the nature of a current wave form which will produce the lowest possible temperature of the junction.

F. *The existence of an extremal of the current wave form in transient thermoelectric cooling*

We finally enquire into the possibility of the existence of a function of current vs. time which causes the maximum possible transitory temperature depression at the junction. Denoting this unknown function by $Z(\Theta)$ we assume again for simplicity as in the previous Subsection that the Peltier coefficient P is a constant independent of temperature. Then the temperature drop of the junction caused by the Peltier effect may be considered to be a superposition of delta-function responses of the type specified by equation (10) and is therefore given by

$$v(0,t)_P = \int_0^{t_1} [-Z(\Theta)\cdot\beta\cdot(t-\Theta)^{-1/2}]d\Theta \tag{69}$$

Expressed otherwise, $v(0,t)_P$ is the convolution of the functions $Z(t)$ and $[-\beta t^{-1/2}]$ where t is the

time of observation ($t > t_1$), t_1 is the time the current flow is terminated ($t_1 > 0$) and

$$\beta t^{-1/2} = G(aK)^{-1}(\kappa/\pi)^{1/2}t^{-1/2} \tag{70}$$

is the temperature due to a single delta function of a strength appropriate to unit A-sec. The factor G in (70) must therefore be defined by

$$G = \tfrac{1}{2}P/4{\cdot}2 \tag{71}$$

During the same time interval extending from 0 to t_1 the Joule heating causes a rise in the junction temperature amounting to

$$v(0,t)_J = \int_0^{t_1} \alpha[Z(\Theta)]^2 d\Theta \tag{72}$$

where $\alpha = 1/4.2a^2\sigma\rho c$.
The junction temperature due to the combined Peltier and Joule effects is therefore given by

$$V(0,t) = v(0,t)_J + v(0,t)_P$$

$$= \int_0^{t_1} \{\alpha[Z(\Theta)]^2 - \beta Z(\Theta)(t-\Theta)^{-1/2}\} d\Theta \tag{73}$$

We now enquire into the form of that function $Z(\Theta)$ which makes $V(0,t)$ a minimum. This is seen to be a simple type of variational problem, because, denoting the integrand of (73) by i, the Euler equation associated with the problem reduces to

$$\frac{\partial i}{\partial Z} = 0 \tag{74}$$

Consequently the wanted function is

$$Z(\Theta)_{ex} = (\beta/2\alpha)(t-\Theta)^{-1/2} \tag{75}$$

where the subscript "ex" indicates that $Z(\Theta)$ is an extremal in the sense of the calculus of variations.†

† Professor R. C. T. SMITH drew our attention to the fact that the same result may be obtained without reference to Euler's equation by completing the square in the integrand of (73) and determining Z from the inequality

$$\int_0^{t_1} \alpha[Z-(\beta/\alpha)(t-\Theta)^{-1/2}]^2 d\Theta - (\beta^2/\alpha)\int_0^{t_1}(t-\Theta)^{-1}d\Theta$$

$$\geqq -(\beta^2/\alpha)\int_0^{t_1}(t-\Theta)^{-1}d\Theta.$$

Using the equality sign and solving for Z gives (75).

The current wave form represented by (75) may still result in either a maximum or a minimum of $V(0,t)$ but it will be seen presently that it actually causes a minimum.

Inserting (75) into (73), the temperature of the junction when measured at an instant t is found to be

$$V(0,t)_{ex} = \int_0^{t_1} \{(\beta^2/4\alpha)(t-\Theta)^{-1} - (\beta^2/2\alpha)(t-\Theta)^{-1}\} d\Theta$$

$$= (\beta^2/4\alpha)\ln[(t-t_1)/t] \tag{76}$$

If we let $t \to +t_1$, i.e. if the time of observation coincides with the instant t_1 when the current $Z(\Theta)$ terminates, the junction temperature tends to

$$\lim_{t \to +t_1} V(0,t)_{ex} = (\beta^2 4\alpha)\ln 0 = -\infty \tag{77}$$

provided $Z(t_1)_{ex}$ in (75) is allowed to rise to infinity.‡

We are thus led to the conclusion that under the conditions set out in Section 1 of this paper there exists for every prescribed time of observation a unique current wave form, more specifically, a continuous function of time that will generate the lowest possible temperature at the junction. Moreover, if the instant of observation coincides with the instant the current terminates, the transient temperature of the junction would actually approach the absolute zero of temperature if the process were not counteracted by changes in σ, K and P which are required to occur by the Third Law of Thermodynamics. In the absence of a detailed calculation of these effects we are able to predict only that much lower temperatures may be expected to be observed with currents varying

‡ It was pointed out by the referee of this paper that a current of the form $I = I_0 \exp(t)$ originally (and erroneously!) believed to cause low-temperature transients does not in fact produce lower transients than a rectangular current pulse.

according to equation (75) than with rectangular current pulses.

These conclusions are also valid for thermocouples of finite dimensions provided the time interval from 0 to t_1 is chosen short enough. It will be noted that the method of measurement described in the following Section allows us to observe transient temperatures within a few milliseconds after the cessation of the current and these effects are therefore the object of a discussion in that section.

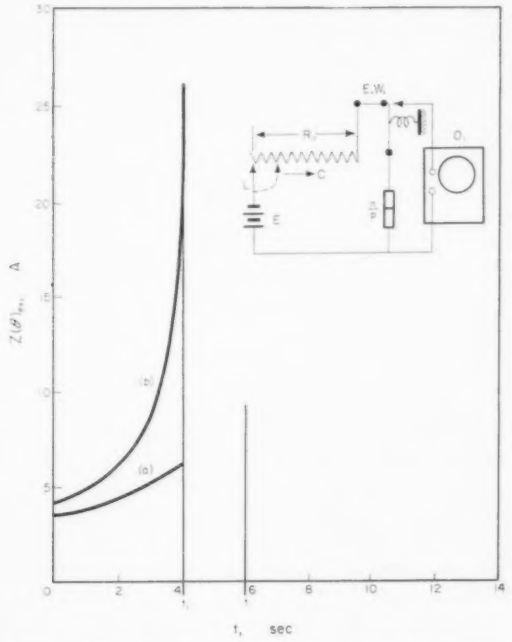


FIG. 5. The function $Z(\Theta)_{ex}$. Time of observation: (a) $t > t_1$, (b) $t \to +t_1$.

Insert: Schematic diagram of a circuit capable of generating the function $Z(\Theta)_{ex}$. [Curve (b) is produced when the slider L moves from left to right with velocity $c \propto 1/\sqrt{t_1 - \Theta}$ along the rheostat of total resistance $R_0 = E(2\alpha/\beta)\sqrt{t_1}$. The other symbols correspond to the circuit elements of Fig. 6.]

The function $Z(\Theta)_{ex}$ is represented in Fig. 5 for both the conditions $t > t_1$ and $t = t_1$ together with a schematic diagram of a circuit suitable for generating such a function.

It is important to note that, after the passage of the current pulse, the temperature of the branches at a large distance from the junction is numerically equal to the junction temperature but of opposite sign.

3. MEASUREMENT OF FAST TRANSIENTS OF TEMPERATURE IN THERMOCOUPLES

A. *The experimental arrangement*

If fast thermal transients of the type considered in Section 2 are to be observed, it is necessary to use a recording instrument with a bandwidth extending from zero to at least several hundred cycles. The technique employed is most conveniently described with the aid of the diagram, Fig. 6. In this diagram, C is the arm of a mechanical changeover switch which allows the thermojunction n/p to be connected to a battery B (a large capacity lead accumulator) or alternatively to the input of an amplifier A. The time required for the switching operation for this type of switch (a modified microswitch) is of the order of a few milliseconds. The amplifier A was designed for this experiment and has been described elsewhere.[16] Briefly, the principle of operation of this amplifier is to apply the unknown potential as a bias voltage to a two-button carbon microphone which is continuously being excited by a constant audio note of frequency 10 kc/s.

It has been shown that the amplified output of the microphone is a linear function of its bias and that, by adjustment of the contact pressure of the carbon buttons, the impedance offered to the source may be set, if necessary, at a value of less than 1Ω. This allows a good match to the internal resistance of the thermojunction to be achieved. The overall sensitivity of this arrangement is more than adequate for our purpose. Long-term stability is difficult to maintain within limits of a few per cent but, since the duration of our experiments rarely exceeds 1 min, and is generally much shorter, stability presents no special problem. To produce short high-current pulses the manual switch $M.S._1$ is first closed. The timing is then left to an exploding wire $E.W.$ which is replaced after each cycle of operation. The strength and duration of the pulse are almost entirely functions of the battery voltage and the diameter of $E.W.$, within a certain range of energy content of the pulse. In the course of these experiments we have acquired a store of data which allows us to adjust the strength and duration of

these pulses as required.† The contact arm C is held against the stop O by the exploding wire $E.W.$ and a spring S_1. (Stop O also functions as a contact. The purpose of this contact is explained below). When switch $M.S._1$ is closed, $E.W.$ fuses after sending a current pulse through the junction, and switch arm C, under the action of spring S_2, closes immediately the right-hand contact connecting the couple to amplifier A and oscilloscope O_1. The current-pulse waveform is recorded on oscilloscope O_2 after amplifying the potential developed by the current across a resistor r which has a resistance of a few milliohms. Fig. 7(a) is a record of such a pulse produced by an exploding wire. The pulse shape is not quite rectangular but shows a slightly sloping top owing to the change of resistance during the heating-up period of the wire. The thermoelectric e.m.f. (Seebeck potential) recorded by oscilloscope O_1 as a function of the time elapsed since the passage of the current pulse can be directly transformed into the temperature of the junction with the aid of the calibration chart, Fig. 8.

Fig. 7(b) is a sample oscillogram of the Seebeck potential of an n- and p-type bismuth telluride junction. Facilities are provided in the output to rectify the 10 kc/s wave form but we found it more convenient to record the unrectified modulated output. The junction temperature then corresponds to the envelope of the modulated output wave. A separate polarity test establishes whether the temperature recorded corresponds to "heating" or "cooling". In all experiments we concentrated on Peltier cooling. Peltier heating was only observed incidentally. The temperature of the hot junction was kept at a constant value, near room temperature, by circulating water. The amplifier was usually calibrated before and after each recording, and the stable operation of the junction was frequently tested by recording the maximum cooling with optimum stationary current.

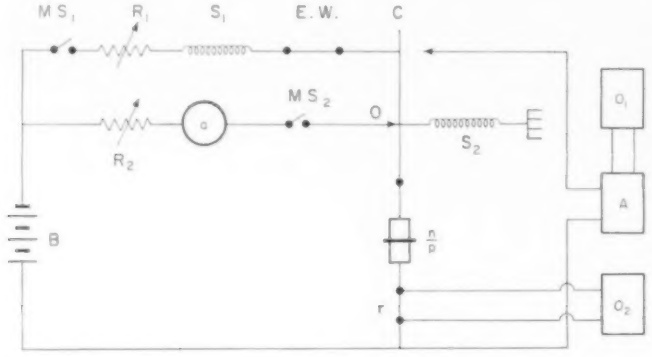


FIG. 6. Schematic circuit diagram of the pulse-producing and measuring circuit.

The current pulse may also be superimposed on a steady current through the junction. With the manual switch $M.S._1$ open and $M.S._2$ closed, a steady current flows through the couple via the series resistance R_2 and ammeter A. When switch $M.S._1$ is closed, $E.W.$ fuses, superimposing a current pulse on to the stationary current and thereafter opening both current paths simultaneously. This is again followed immediately by the beginning of the recording of the Seebeck potential. In all cases "long" pulses of one to several seconds duration and of comparatively low current (about 10A) are obtained by operating

† With the lower-current pulses the wires do not actually explode in the manner which has given the name to this effect. They simply heat up and then fuse in the middle. However, for purposes of timing it does not seem to matter whether the circuit is broken by radial outward diffusion or by fusion of the material of the wire.

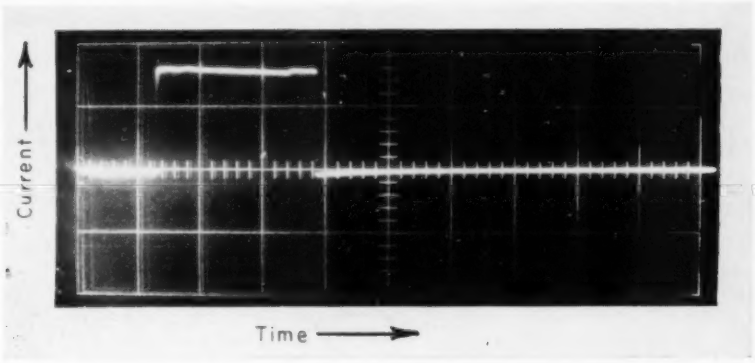


FIG. 7(a). Oscillogram of a current pulse ($I = 77$ A, $\tau = 20$ msec) produced by an exploding wire.

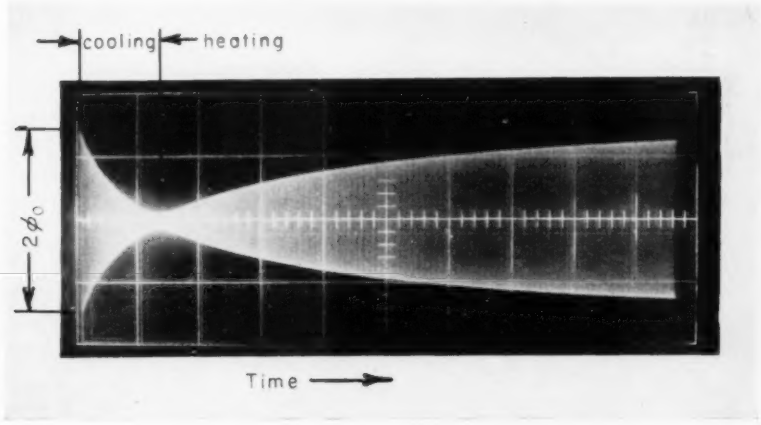


FIG. 7(b). Oscillogram of the Seebeck potential developed across the thermocouple after the passage of a current pulse.

[facing p. 254

the switch arm C manually and using a stopwatch.

The time resolution of the whole arrangement, i.e., the time elapsed between the moment the current ceases and the moment the recorded temperature is readable from the oscilloscope screen, is mainly governed by the inertia of the contact arm C, and we found it difficult to reduce this "dead time" interval to much less than 10 msec. Electronic switching, though feasible, would make the whole arrangement much more complicated. Though there exists this slight uncertainty about the loss of time in measuring the thermoelectric potential, this disadvantage is outweighed by the fact that, by the nature of the experiment, there is never any doubt about the location of "the junction". This difficulty is always present in methods using external thermometry.

B. *Measuring procedure and calibration*

The measuring circuit allows the passing of current pulses of various strength and duration through the junction and the oscillographic recording of the Seebeck potential immediately after the instant that the current passes. If the switching time is negligible compared to the decay time of the transient, we measure, in the notation of Section 2, $\phi(t)$ for $t > \tau$. If this criterion is not applicable, the peak value of the record supplies only a lower limit to the Seebeck potential. The peak value of the potential was read from 35-mm film records by means of a travelling microscope. Since the potential curve is sometimes very steep at this point [cf. Fig. 7(b)] the maximum potential could not always be read unambiguously and therefore several records were usually taken under identical conditions. Even then the experimental points sometimes showed a considerable scatter. However, as seen from Figs. 9–11, the general tendency can always be discerned.

The Seebeck potential must be converted into temperature. For this purpose we have used published results of measurements of the Seebeck coefficient together with measurements of this coefficient on our own samples at room temperature.

(a) *Temperature region* 150–300°K. Reference has already been made to the result[11] that the Seebeck coefficient varies logarithmically in this region. Using equation (40) of Section (2), the Seebeck potential is then given, by integration, as

$$\phi = 2[r(T\ln T - T) + C_0 T]_T^{T_0} \tag{78}$$

where r is a mean value of the slope of the "$|S_{n,p}|$vs.ln$T$" curve for the n- and p-type materials taken singly, and where the constant C_0 is to be determined from the requirement that $|S_n| = |S_p| = 175\mu$V/deg, the average value of our n- and p-type materials at room temperature $T_0 = 290$°K. Adopting Goldsmid's average value for r of 0·124 mV/deg (corresponding to a curve with an S value of 175 μV/deg at room temperature) a calibration curve is obtained whose lowest temperature of 150°K corresponds to a Seebeck potential of $\sim$ 39 mV.

(b) *Temperature region* 6–150°K. For this region also, measurements of the Seebeck coefficient of (monocrystalline) bismuth telluride are available.[17] These measurements indicate that it is sufficient for our purpose to assume that the Seebeck coefficient is proportional to the absolute temperature between 6 and 150°K. This is because for low transient temperatures our measurements become progressively less accurate. At 150°K, where the two regions meet, equation (40) yields a value of $S \simeq 93\mu V$/deg. The factor of proportionality is therefore $\simeq 93/150 = 0{\cdot}62$ for our specimen. Then we have for the lower region

$$d\phi = 2 \times 0{\cdot}62 \times 10^{-3} T dT \tag{79}$$

and on integrating this expression

$$\begin{aligned} \phi_{(150)} &= [0{\cdot}62 \times 10^{-3} T^2]_T^{150} \\ &= 14 - 0{\cdot}62 + 10^{-3} T^2 \text{ mV} \end{aligned} \tag{80}$$

where $\phi_{(150)}$ is the Seebeck potential when the cold junction is at T°K and the hot junction at 150°K. Over the composite region from 6°K to room temperature, T_0, the Seebeck potential is represented by

$$\phi = \int_T^{T_0} S dT = \int_{150}^{T_0} S dT + \int_T^{150} S dT \tag{81}$$

so that in the lower region and when the cold junction is at T°K and the hot junction at room

temperature T_0

$$\phi_{(T_0)} = 39+14-0{\cdot}62\times10^{-3}T^2$$
$$= 53-0{\cdot}62\times10^{-3}T^2\,\text{mV} \qquad (82)$$

Relations (78) and (82) are both plotted in the composite calibration chart, Fig. 8, which we have used to convert the measured Seebeck potential into temperature and to plot the experimental curves Figs. 9–11.

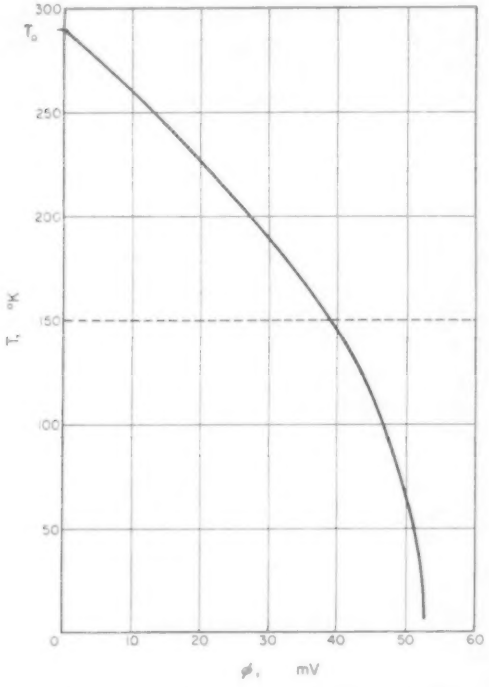


FIG. 8. Composite calibration chart, $T = f(\phi)$, for the temperature region 6–290°K (room temperature).

Regarding the method of calibration described, it may be surmised that a direct calibration of the individual couples used would have been a more satisfactory procedure. However, there are at present no low-temperature facilities available in the authors' laboratory. Furthermore, it is almost certain that the advantage of a direct calibration would be lost by the necessity to unsolder the *n*- and *p*-branches again before or after calibration. In order to make the couples correspond as closely as possible to our theoretical model, the branches were joined with the aid of an ultrasonic soldering iron using an absolute minimum of solder. Therefore the couples cannot be calibrated *in situ* but must be unsoldered and joined to heat sources and sinks, whereby a disturbance of the thermal and electrical properties and of the geometry can hardly be avoided.

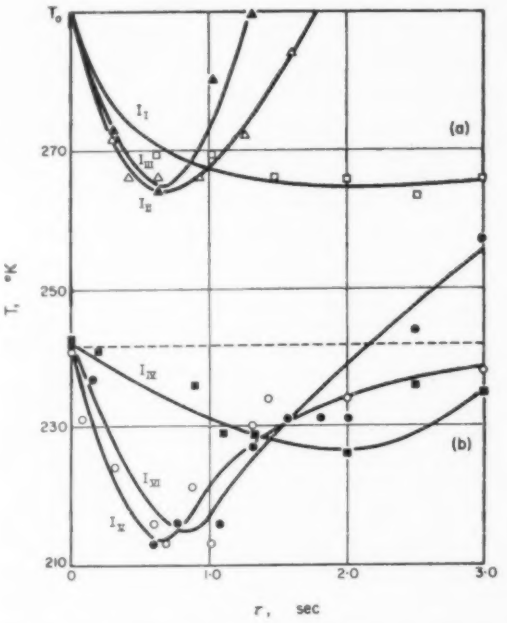


FIG. 9(a). Couple No. 6. The junction temperature measured as a function of time after the passage of current pulses of strength $I_1 = 16$ A, $I_2 = 28$ A, $I_3 = 32$ A.

FIG. 9(b). Couple No. 6. The junction temperature measured as a function of time after the passage of current pulses of strength $I_4 = 9$ A, $I_5 = 21$ A, $I_6 = 33$ A, superimposed on an optimum steady current of 8 A. The dotted line corresponds to the optimum steady temperature −31°C.

C. *Experimental results*

The experimental results are best displayed in the form of families of curves of $V(0,\tau)$ vs. τ for different values of the pulse current I, and curves of $V(0,\tau)$ vs. $\zeta = I\sqrt{\tau}$ for different values of the pulse current. Amongst a number of thermocouples made up, two specimens, proved to be particularly convenient and the results for these couples are given here. At room temperature these

Table 1. *Dimensions of thermocouples and average thermoelectric properties of bismuth telluride*

Cross-sectional area of couple No. 6	a	0·16 cm^2
Cross-sectional area of couple No. 7	a	0·04 cm^2
Length of couple No. 6	l	2 cm
Length of couple No. 7	l	0·8 cm
Bismuth Telluride		
Average density of n- and p-type	ρ	8 g/cm^3
Average specific heat of n- and p-type	c	0·13 cal/g-deg
Average heat conductivity of n- and p-type	K	$4{\cdot}8\times10^{-3}$ cal/sec-cm-deg
Average electr. conductivity of n- and p-type	σ	1000 Ω^{-1} cm^{-1}
Average diffusivity of n- and p-type	κ	$4{\cdot}6\times10^{-3}$ cm^{2}sec^{-1}
Average Seebeck coefficient of n- and p-type	S	$1{\cdot}75\times10^{-4}$ V/deg

couples (No. 6 and 7) had a Seebeck coefficient of 350 μV/deg when mounted, which corresponds to the middle of the range of curves in Refs. 11 and 17. In addition, the figure of merit of both samples was comparatively high. The dimensions and average properties of the two couples are summarized in Table 1.

Fig. 9(a) shows $V(0,\tau)$ with I as parameter for couple No. 6 when cooling started from room temperature. The maximum cooling was about 26°C and was the same for the three values of current shown. This is very close to the value 28°C calculated in Section 2C (cf. curve c, Fig. 3) under the assumption that S is a constant, and demonstrates that even for such comparatively small temperature differences the assumption that P is a constant is not adequate. On the other hand, the minima occur for a value of ζ of about 22 A-sec$^{1/2}$ as compared to 48 as required by the theory [see Fig. 10(a)]. We comment on this discrepancy later in this Section.

The temperature $V(0,\tau)$ for the same couple is displayed in Fig. 9(b) when the current pulse was superimposed on the stationary optimum current of 8 A. In this case an additional 29°C of cooling below the stationary cooling of about 50°C was observed. This is seen to be in agreement with the discussion in Section 2E. Deviations were again observed for large values of τ (compare the curve for the current pulse of 9 A). As shown in Fig. 11(b) for pulses of 9 and 21 A, the minima occur for substantially the same value of ζ ($\sim$ 22 Asec$^{1/2}$) as without stationary cooling.

These results indicate that for the longer pulses and lower currents the couple can no longer be considered to be a "long" couple and that the proximity of the hot junction makes itself felt. This is also borne out by the observation that, for the pulse of 33 A superimposed on a steady current of 8 A, ζ_{min} is about 35 Asec$^{1/2}$ [see Fig. 11(b)] which is much closer to the theoretical value of 48 Asec$^{1/2}$ than the ζ_{min} for the lower current pulses.

The results for couple No. 7 which has a cross-section only about one quarter of couple No. 6 require a more detailed discussion. Even in the early stages of our experiments it appeared that, in particular for couples with small cross-sectional area, the minimum of the temperature was considerably lower than one would expect from the simple relationships of Section 2A–D. The curves for couple No. 7 [Fig. 10(a, b)] show that the maximum cooling increases strongly with increasing current, and very low transient temperatures indeed are seen to occur. Fig. 10(a) shows $V(0,\tau)$ with I as parameter when cooling started from room temperature. Even for a comparatively low current (8 A), cooling greater than the maximum cooling for the optimum steady current of 5 A was obtained. With high currents of more than 20 A

a few singular readings of the Seebeck potential were in fact close to or even slightly exceeding the limit of the calibration chart of ~ 53 mV.

Although the experimental error in this region is admittedly very high, it seems without doubt that transient temperatures below 100°K have occurred. If the pulse currents are increased further still, the magnitude of the Seebeck potential

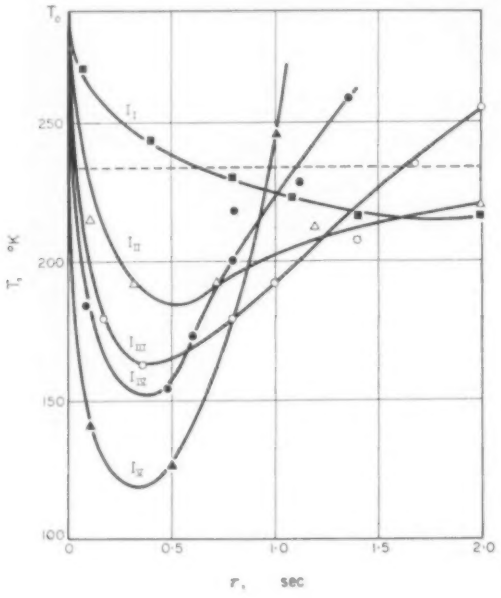


FIG. 10(a). Couple No. 7. The junction temperature measured as a function of time after the passage of current pulses of strength $I_1 = 8$ A, $I_2 = 12$ A, $I_3 = 16$ A, $I_4 = 20$ A, $I_5 = 22$ A.

diminishes again, but there are indications that this is due to the finite switching time. When the current pulses are superimposed on a stationary optimum current, a similar behaviour is observed for current pulses greater than 5 A. For pulses neither with nor without stationary cooling can the observed Seebeck potentials be reconciled with the calculations of Section 2A–D by assuming unusually large values of S or abnormal temperature dependence of S for this couple. Furthermore, we shall now show that *with bismuth telluride couples* the enhanced cooling observed cannot be attributed to the Thomson effect, although the contribution of this effect becomes more effective the shorter the transient currents are, as was pointed out in Section 2D.

The Thomson coefficient of our samples was not measured, but, using published data[4] of this coefficient for Bi_2Te_3 ($\gamma = 1{\cdot}5 \times 10^{-4}$ V/deg), we calculate the temperature drop due to the combined Peltier and Thomson effect for the extreme case of a delta-function-type of current pulse

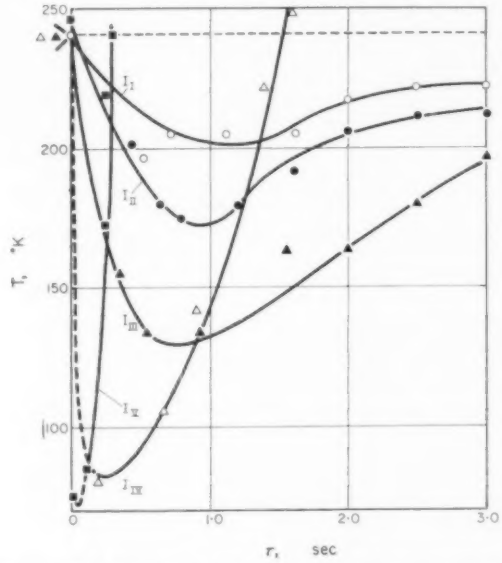


FIG. 10(b). Couple No. 7. The junction temperature measured as a function of time after the passage of current pulses of strength $I_1 = 3$ A, $I_2 = 5$ A, $I_3 = 9$ A, $I_4 = 20$ A, $I_5 = 100$ A, superimposed on an optimum steady current of 5 A. The dotted line corresponds to the optimum steady temperature of -32°C.

where the two effects by necessity add algebraically. Rewriting equation (19), we have in first approximation

$$V_{\min} = -\frac{1}{16}\frac{\pi^{-1}}{4{\cdot}2}\frac{\sigma}{K}(P+\gamma V_{\min})^2 \qquad (83)$$

and neglecting the small term $(\gamma V_{\min})^2$

$$V_{\min} \simeq -\frac{1}{16}\frac{\pi^{-1}}{4{\cdot}2}\frac{\sigma}{K}\frac{P^2}{(1-2\gamma P)} = -\frac{V_{\min}(\text{Peltier})}{(1-2\gamma P)} \qquad (84)$$

Using the above value of γ and $P \simeq 1 \times 10^{-1}$ from Table 1, we see that the denominator of (84) has a numerical value of $\sim (1 - 1{\cdot}5 \times 10^{-5})$ near room temperature, so that a change in $V_{\min}$ would not be detected with our method.

However, in the course of these experiments it became evident that the wave form of the current pulse had a decisive influence on the magnitude of the observed transients, and this observation led us to the calculations given in Section 2F. The results of these calculations allow us to explain all experimental results if for this purpose we divide the pulse durations τ, for a couple of given cross-section a, into four characteristic regimes:

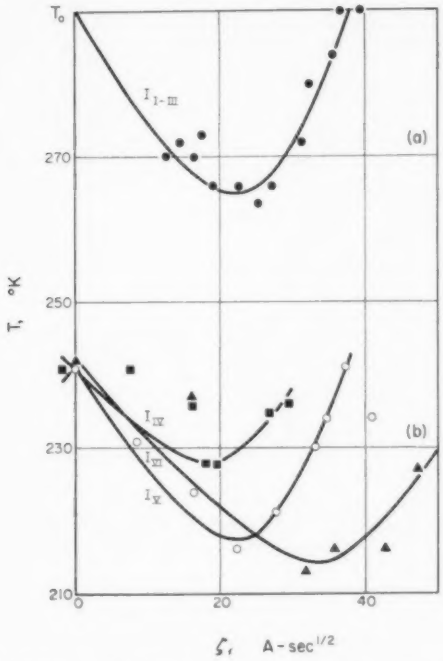


FIG. 11(a). Couple No. 6. The temperature measured at the junction as a function of the parameter $\zeta = I\sqrt{\tau}$. The currents I_1, I_2, I_3 are the same as those used in Fig. 9(a).

FIG. 11(b). Couple No. 6. The temperature measured at the junction as a function of the parameter $\zeta = I\sqrt{\tau}$. The currents I_4, I_5, I_6 are the same as those used in Fig. 9(b). The pulses are superimposed on an optimum steady current of 5 A.

(a) The time interval τ is so long (for example $\tau > 2$ sec for couple No. 7) that we no longer have $l \gg \sqrt{2\kappa t}$. Then the effect of the hot junction is already noticeable, i.e. the temperature minima are not as low as expected and $\zeta_{\min}$ is shifted towards smaller values [cf. curve I_1, Fig. 9(a)].

(b) The time interval is of an intermediate length (for example $1 < \tau < 2$ sec for couple No. 7). Then the behaviour corresponds to the one calculated for a finite rectangular pulse [cf. curve I_2, Fig. 9(a)].

(c) A region of very short time intervals ($\tau < 1$ sec for couple No. 7). In this range the deviation of the wave form from a rectangular shape becomes important. In particular it was found that whenever the instantaneous rise of the current was delayed by the inclusion of large wire-wound resistances with an appreciable inherent inductance, then very low but short-lived temperature transients were observed [cf. curves I_3, I_4, Fig. 10(b)]. It is not difficult to calculate the effect of a current wave form which is not an extremal as $Z(\Theta)_{ex}$ in Section 2F, but is "in the neighbourhood" of an extremal. The residual circuit inductance was actually measured and it was found that the time delay was of an order of magnitude that could well account for the temperatures being produced by the mechanism analysed in Section 2F. In addition, experiments were performed in which the function $Z(\Theta)_{ex}$ was generated as faithfully as possible with the aid of the circuit shown schematically in Fig. 5. It was found that short transient temperatures below 100°K could usually be produced by this method in particular when the junction was precooled by a steady current. The exactness of simulating the wave form is reasonably critical. If the generated function deviates too much from $Z(\Theta)_{ex}$ the thermocouple is burnt out at high currents.

(d) If τ is still shorter ($< 0{\cdot}1$ sec for couple No. 7) and the currents correspondingly high, observation of the temperature minimum becomes increasingly difficult because of the shortness of the transient temperature and the time resolution of the changeover switch [cf. curve I_5, Fig. 10(b)].

Detailed measurements are represented for couple No. 7 in Fig. 10(a, b), corresponding respectively to Fig. 9(a, b) for the larger couple No. 6. For intermediate values of current maximum

cooling occurs for the same value of ζ_{min} (9 A-sec$^{1/2}$). These diagrams well illustrate the four different regimes of behaviour for very high and very low currents.

4. CONCLUSION

It has been pointed out by STIL'BANS and FEDOROVICH[3] that the occurrence of temperature transients may be turned to advantage in some instruments utilizing the thermoelectric effect, as for example in hygrometers. Although this is beyond the scope of this paper, one may speculate on whether transient low temperatures such as have been discussed in the previous sections will find practical applications in producing low temperatures. It is clear that refrigerating devices based on such a principle would involve very awkward thermal and electrical moving contacts. There is, however, one possibility that comes to mind and in which these difficulties may not be unsurmountable. If the nozzles and connecting tubes in a Joule–Thomson type of refrigerator were cooled thermoelectrically, it may be possible to obtain a considerable advantage by pulsing the gas stream in this refrigerator in proper phase relation to the electric current. In this laboratory, the interest is in certain solid-state aspects of these transients rather than in their applications.

Acknowledgements—We are greatly indebted to Professor J. M. SOMERVILLE, Professor of Physics in the University of New England, for advice and encouragement with this work. Professor R. C. T. SMITH, Professor of Mathematics in this University, kindly read the draft of the revised Section 2F and pointed out an alternative derivation of the main result of that Section. Thanks are due to Dr. P. E. FIELDING of the Department of Physical Chemistry of this University for advice in the preparation of samples and for the loan of some of his own ingots. The work has greatly benefited from discussions on various aspects of this research we have had with Drs. F. H. HIBBERD and N. H. FLETCHER of this Department. Mr. M. IDNURM and Mr. P. T. BAILEY assisted us with experiments and we have been ably assisted with design and construction work by Mr. D. A. ATKINSON, the Laboratory and Workshop Manager of this Department, and by the workshop staff.

We are particularly grateful to the Referee of this paper for his most competent criticism which enabled us to eliminate some serious errors in the manuscript.

REFERENCES

1. W. C. WHITE, *Elect. Engng., N.Y.* **70,** 589 (1951).
2. A. H. BOERDIJK, *J. Appl. Phys.* **30,** 1080 (1959).
3. L. S. STIL'BANS and N. A. FEDOROVICH, *Soviet Phys., Tech. Phys.* **3,** 460 (1958).
4. J. E. PARROTT, *Solid-State Electron.* **1,** 135 (1960).
5. P. E. GRAY, *The Dynamic Behaviour of Thermoelectric Devices.* John Wiley, New York and London (1960).
6. C. D. HODGMAN, *Handbook of Chemistry and Physics* (41st ed.) p. 291 ff. Chemical Rubber Publishing Co., Cleveland (1959–60).
7. A. ERDÉLYI, *Tables of Integral Transforms*, p. 125 ff. McGraw-Hill New York (1954).
8. H. S. CARSLAW and J. C. JAEGER, *Conduction of Heat in Solids*, p. 494 ff. Oxford University Press (1959).
9. K. ALTENKIRCH, *Phys. Z.* **12,** 920 (1911).
10. M. B. GRIER, *Proc. I.R.E.* **47,** 1515 (1959).
11. H. J. GOLDSMID, *Proc. Phys. Soc. Lond.* **71,** 633 (1958).
12. D. W. JORDAN, *Brit. J. Appl. Phys.* **12,** 14 (1961).
13. A. F. IOFFE, *Semiconductor Thermoelements and Thermoelectric Cooling.* Infosearch, London (1957).
14. H. J. GOLDSMID, *Applications of Thermo-Electricity* Methuen, London and New York (1960).
15. R. V. CHURCHILL, *Operational Mathematics*, (2nd ed.) McGraw-Hill New York, (1958).
16. K. LANDECKER, *J. Electron. Control* **3,** 218 (1957).
17. P. A. WALKER, *Proc. Phys. Soc., Lond.* **76,** 113 (1960).

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 261–267. Printed in Great Britain

IMPROVEMENT OF SEMICONDUCTING DEVICES BY ELASTIC STRAIN

W. G. PFANN

Bell Telephone Laboratories, Inc., Murray Hill, New Jersey

(*Received* 16 *June* 1961; *in revised form* 31 *July* 1961)

Abstract—The mobilities of current carriers in germanium, silicon and other semiconductors can be changed by ~ 50 per cent or more by elastic strain. The changes in mobility can be positive or negative, and they can be produced in selected directions in the crystal. One can, therefore, improve, or compensate against temperature change, certain electrical properties of junction devices, such as transistors, varactor diodes, tunnel diodes, solar diodes, and perhaps even thermoelectric devices. Some of these possibilities are examined in semi-quantitative fashion.

Résumé—Les mobilités de porteurs de courant dans le germanium et le silicium peuvent être changées de 50 pour cent ou plus par déformation élastique. Ces changements de mobilité peuvent être positifs ou negatifs et peuvent être produits dans certaines directions choisies du cristal. On peut donc améliorer ou compenser en fonction des changements de température certaines propriétés électriques de certains éléments à jonction tels que les transistors, les doides varactor, les diode tunnel, les diodes solaires, et même certains éléments thermoélectriques. Quelques-unes de ces possibilités sont examinées de façon semi-quantitative.

Zusammenfassung—Die Beweglichkeit der Stromträger in Germanium, Silizium und anderen Halbleitern kann durch elastische Spanungskräfte um etwa 50 Prozent oder mehr verändert werden. Die Änderungen der Beweglichkeit können positiv oder negativ sein und können im Kristall in ausgewählten Richtungen erzeugt werden. Auf diese Weise lassen sich also Temperaturveränderungen, gewisse elektrische Eigenschaften von Geräten mit Übergängen (Transistoren, Varactor-Dioden, Tunnel-Dioden, Sonnen-Dioden) und vielleicht sogar von thermoelektrischen Geräten verbessern oder kompensieren. Einige dieser Möglichkeiten werden auf quasi-quantitative Weise untersucht.

1. INTRODUCTION

UNCONTROLLED elastic stresses are generally regarded as a nuisance in the fabrication of semiconducting devices. If large enough, they may lead to fracture, or at least plastic deformation, of the semiconductor, with attendant impairment of the properties and uniformity of the devices. Without denying this, we take the opposite view in this paper, namely that the judicious application of controlled elastic stress can markedly improve some of the important properties of basic devices, such as transistors and diodes.

Although it has been recognized that the mobilities of the carriers in germanium and silicon are very sensitive to elastic stress (piezoresistive effect) and that the resistance of a p–n junction is sensitive to hydrostatic pressure (mainly through the energy-gap effect), the idea of using such effects to make better devices does not appear to have received the attention it deserves. This lack of attention may be in part because the strains required are large, lying between $\sim 10^{-3}$ at which these materials commonly fracture, and $\sim 10^{-2}$, the highest elastic strain thus far achieved in them, and in part because the effects of large elastic strains on the behavior of p–n junction devices are rather complicated and not widely appreciated.

In this paper we assume that strains $> \sim 10^{-3}$ can be produced (although certain applications permit smaller strains), and we suggest briefly some ways of doing so. We assume that changes in mobility of $\sim \pm 50$ per cent, in selected crystal directions, and changes in energy gap of several kT at room temperature, can be produced by the application of strains of the magnitude indicated. Examples are given of how elastic stresses may improve the current multiplication factor and

gain-band product of a transistor, or the *RC* product of a varactor diode.

A corollary to these ideas, suggested by I. M. Ross and J. M. Early, is that part of the scatter in properties of junction devices now in manufacture may stem from uncontrolled elastic stresses. It is hoped that the semiquantitative treatment offered here will lead to a better understanding of such effects.

2. PIEZO-RESISTANCE RELATIONS

The mobilities of holes and electrons in germanium or silicon in any given direction of the crystal depend only slightly on hydrostatic stress. However, they can be extremely sensitive to uniaxial stress, and this sensitivity will be expressed here in terms of the longitudinal and transverse piezo-resistance constants, π'_{11} and π'_{21}, respectively. The use of primes denotes that the coefficients refer to arbitrary directions in the crystal. Thus π'_{11} refers to the resistance change in the 1′ direction in response to a normal stress in the 1′ direction; and $\pi'_{21} = \pi'_{12}$ refers to the resistance change in the 2′ (or 1′) direction in response to a normal stress in the 1′ (or 2′) direction. The effects of shear stress could be considered as well, but for most of the orientations considered here the appropriate shear coefficients are zero, and these will be neglected for simplicity. Expressions for π'_{11} and π'_{21} for cubic crystals are:[1]

$$\pi'_{11} = \pi_{11} - 2A(l_1^2 m_1^2 + l_1^2 n_1^2 + m_1^2 n_1^2) \qquad (1)$$

$$\pi'_{21} = \pi'_{12} = \pi_{12} + A(l_1^2 l_2^2 + m_1^2 m_2^2 + n_1^2 n_2^2) \qquad (2)$$

where

$$A = \pi_{11} - \pi_{12} - \pi_{44} \qquad (3)$$

in which π_{11}, π_{12}, π_{44} are the fundamental piezo-resistance constants[2], and l_1, m_1, n_1 and l_2, m_2, n_2 are the direction cosines of the 1′ and 2′ directions with respect to the cubic axes of the crystal. Table 1 lists expressions for π'_{11} and π'_{12} for various crystal directions of interest. For resistivities $\approx 1\ \Omega$ cm, one of the fundamental coefficients is large, $\sim 10^{-10}$ cm^2/dyne, for each of the four materials considered in detail here, namely *p*- and *n*-germanium, and *p*- and *n*-silicon. For the first three, π_{44} is large, the other coefficients small. For the last, π_{11} is large and $\pi_{12} \sim -\frac{1}{2}\pi_{11}$, π_{44} being small. The coefficients get smaller as the resistance decreases.[3,4] The large coefficients are negative for electrons, positive for holes, assuming a tensile stress.

Table 1. *Longitudinal and transverse piezo-resistance coefficients for various directions in cubic crystals*

Longitudinal direction, 1	Longitudinal, coefficient, π_{11}	Transverse direction, 2	Transverse, coefficient, π_{12}
001	π_{11}	010	π_{12}
001	π_{11}	110	π_{12}
111	$\frac{1}{3}[\pi_{11}+2\pi_{12}+2\pi_{44}]$	$1\bar{1}0$	$\frac{1}{3}[\pi_{11}+2\pi_{12}-\pi_{44}]$
111	$\frac{1}{3}[\pi_{11}+2\pi_{12}+2\pi_{44}]$	$11\bar{2}$	$\frac{1}{3}[\pi_{11}+2\pi_{12}-\pi_{44}]$
$1\bar{1}0$	$\frac{1}{2}[\pi_{11}+\pi_{12}+\pi_{44}]$	111	$\frac{1}{3}[\pi_{11}+2\pi_{12}-\pi_{44}]$
$1\bar{1}0$	$\frac{1}{2}[\pi_{11}+\pi_{12}+\pi_{44}]$	001	π_{12}
$1\bar{1}0$	$\frac{1}{2}[\pi_{11}+\pi_{12}+\pi_{44}]$	110	$\frac{1}{2}[\pi_{11}+\pi_{12}-\pi_{44}]$
$1\bar{1}0$	$\frac{1}{2}[\pi_{11}+\pi_{12}+\pi_{44}]$	$11\bar{2}$	$\frac{1}{6}[\pi_{11}+5\pi_{12}-\pi_{44}]$
$11\bar{2}$	$\frac{1}{2}[\pi_{11}+\pi_{12}+\pi_{44}]$	$1\bar{1}0$	$\frac{1}{6}[\pi_{11}+5\pi_{12}-\pi_{44}]$
$1\bar{1}0$	$\frac{1}{2}[\pi_{11}+\pi_{12}+\pi_{44}]$	$22\bar{1}$	$\frac{1}{6}[4\pi_{11}+5\pi_{12}-4\pi_{44}]$
$22\bar{1}$	$\pi_{11}-\frac{1}{2}\frac{6}{9}[\pi_{11}-\pi_{12}-\pi_{44}]$	$1\bar{1}0$	$\frac{1}{6}[4\pi_{11}+5\pi_{12}-4\pi_{44}]$

The definitions of π'_{11} and π'_{21} are

$$\left(\frac{\Delta\rho}{\rho_0}\right)_{1'} = \pi'_{11}\sigma'_1; \quad \left(\frac{\Delta\rho}{\rho_0}\right)_{2'} = \pi'_{21}\sigma'_1 \qquad (4)$$

where σ_1 denotes a simple tensile stress in the 1′ direction. Thus, a positive π denotes that the mobility *de*creases with tension, *in*creases with compression.

In *n*-silicon, for which the (100) valley model applies, if all electrons could be forced by strain

into one valley, a mobility ratio $\mu_\perp/\mu_\parallel$ of about 5 could be expected in directions normal and parallel to $\langle 100 \rangle$. In n-germanium, for which the (111) valley model applies. if all electrons could be forced into one valley, a ratio $\mu_\perp/\mu_\parallel$ of about 16 could be expected for directions normal and parallel to $\langle 111 \rangle$. To the writer's knowledge, no one has succeeded in going this far at room temperature, but resistance changes of ~ 75 per cent have been produced in silicon. At present there is not a very solid basis for predicting the maximum attainable anisotropy for holes in either material at room temperature.

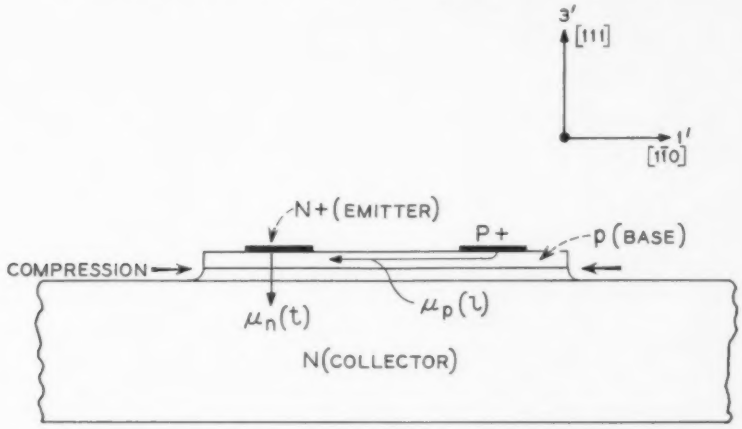


FIG. 1. Germanium n–p–n transistor under compressive stress parallel to plane of junction.

3. ENERGY-GAP EFFECTS

The energy gap is affected by hydrostatic pressure, which changes the density of minority carriers, and hence the saturation current of an ordinary Shockley-type p–n junction. The coefficients $(\delta Eg/\delta P)_T$ are $\sim +5\times10^{-12}$ eV cm^2/dyn for germanium, and $\sim -2\times10^{-12}$ for silicon.[5] A uniaxial compression can be regarded, in this respect, as a hydrostatic pressure P equal to 1/3 the compressive stress.

The effect of a directional or hydrostatic stress on the saturation current I_s of an ideal p–n junction, and hence on the conductance at any given voltage, can be seen from the following equation

$$I_s = I_{ps}+I_{ns} = e\left(p_n\sqrt{\frac{D_p}{\tau_p}}+n_p\sqrt{\frac{D_n}{\tau_n}}\right) \quad (5)$$

where p_n and n_p are the minority-carrier concentrations, the D's are diffusivities (proportional to the mobilities) ,and the τ's are lifetimes. (We assume τ to be independent of elastic strain.) Depending on the material, the orientation of the junction and the direction of the stress (or stresses), the mobility effects and the energy-gap effects can be in the same or opposite sense. For germanium the energy gap effect will usually be larger, although, as shown below, the mobility effects in general are the more important to over-all device behavior.

4. GAIN-BAND PRODUCT OF A TRANSISTOR

A. *Germanium*

A figure of merit F for an n–p–n transistor of optimum design is[6]

$$F \equiv \frac{\mu_n\mu_p}{K^{1/2}} \quad (6)$$

where K denotes dielectric constant. Close study of equation (6) reveals that μ_n, the electron mobility, is present because it determines the transit time of electrons *across* the base layer. Hence large μ_n is wanted in the direction normal to the plane of the p-type base layer. The quantity μ_p is present because it determines the resistance of the base layer to current that flows mainly *along* the base layer from the base contact to the emitter contact. Since the base current consists

mainly of holes, large μ_p is wanted in the plane of the base layer.

It turns out that both of these quantities can be increased substantially in an n–p–n mesa transistor made of germanium in the (111) plane by a compressive stress parallel to the (111) plane, as shown in Fig. 1. The critical mobilities are denoted $\mu_n(t)$ and $\mu_p(l)$ where t and l signify transverse and longitudinal with respect to the base layer.

The stress direction illustrated in Fig. 1 is $[1\bar{1}0]$. For the orientation of Fig. 1, π'_{11}, to which the desired change in $\mu_p(l)$ is proportional, is

$$\pi'_{11} = \tfrac{1}{2}(\pi_{11}+\pi_{12}+\pi_{44}) \simeq +\tfrac{1}{2}\pi_{44} \qquad (7)$$

while, π'_{31}, to which the desired change in $\mu_n(t)$ is proportional, is

$$\pi'_{31} = \tfrac{1}{3}(\pi_{11}+2\pi_{12}-\pi_{44}) \simeq -\tfrac{1}{3}\pi_{44} \qquad (8)$$

Both of these are large, and in the correct sense.

If a second, equal, compressive stress, parallel to $[11\bar{2}]$, is added in the plane of the base layer, the change in $\mu_n(t)$ is doubled, and the change in $\mu_p(l)$ is decreased by $\frac{1}{3}$, leading to a net increase in F. For this stress, π'_{31} is also given by (8), while

$$\pi'_{12} = \tfrac{1}{6}(\pi_{11}+5\pi_{12}-\pi_{44}) \simeq -\tfrac{1}{6}\pi_{44} \qquad (9)$$

The change in $\mu_n(t)$ is then proportional to $\sim -(\frac{2}{3})\pi_{44}$; the change in $\mu_p(l)$ is proportional to $(\pi_{11}+\pi_{12}) \simeq =(\frac{1}{3})\pi_{44}$.

Such a biaxial stress system might be imposed, for example, by the difference in thermal contraction between the germanium wafer and the stud to which it is hard soldered. The uniaxial stress might be applied, for example, by bending the wafer about an axis normal to $[1\bar{1}0]$.

As a matter of fact, the results cited above are valid for any two normal directions, 1′ and 2′, in the (111) plane and hence could be applied to an emitter and base of concentric circular symmetry under radial compressive stress.

Additional benefits of the arrangement just described are that compression increases the energy gap of germanium, thereby reducing the saturation current; very slightly decreases the dielectric constant K; and strongly *de*creases $\mu_p(t)$ the hole mobility normal to the plane of the base layer, which, together with the increase in $\mu_n(t)$, substantially increases the γ-factor of the emitter. Also, $\mu_n(l)$ decreases, which should reduce fringing effects and lead to a greater equalization of electron transit times of emitted electrons.

For a germanium p–n–p transistor the same scheme can be used, except that the stresses parallel to the (111) plane must now be tensile. However, a compressive stress normal to the (111) plane will also increase $\mu_n(l)$, $\mu_p(t)$ and E_g, and decrease $\mu_n(t)$.

B. *Silicon*

Silicon is considerably more complicated than germanium because electrons obey the $\langle 100 \rangle$ valley model while holes have the symmetry of a $\langle 111 \rangle$ valley model. If the current is in a direction symmetrical to the particular valley direction, the mobility change will be negligibly small. Nevertheless, there are various useful combinations, and examples are given in Table 2, with qualitative notes on their respective merits. The directions and symbols refer to Fig. 1.

In Table 2, the two mobility changes that are in adjacent columns enter into the expression for F in equation (6); the third mobility change affects the emitter γ-factor. In this way the reader can readily tell which factor, F or γ, is referred to by the notation "Good" or "Bad". In the column headed ΔE_g, "Good" signifies an increase because this would decrease the saturation current I_s. However, it is quite possible that the mobility changes can in certain cases offset the effect on I_s of the indicated change in E_g.

5. *RC* PRODUCT OF A VARACTOR DIODE

In a typical varactor diode, a very heavily doped p-type alloy or diffusion region forms a p–n junction with a less heavily doped n-type base material. The R in the RC product is determined mainly by the resistance of the base material. Finding the smallest value of RC involves a compromise between the reduction in R and the increase in barrier capacitance C upon increase in the donor concentration of the base material. The decrease in R predominates as the donor concentration increases, until a concentration is reached at which impurity scattering significantly lowers the mobility of electrons. So an optimum donor concentration exists.

By applying elastic strain in the proper direction, the optimum donor concentration can be increased and hence the RC product can be lowered.

For an n-germanium base layer, with the junction lying in the (111) plane, as shown in Fig. 2, a radial compressive stress σ_r in the (111) plane will produce the desired effect, the relative resistance change for electrons travelling normal to the junction being

$$\left(\frac{\Delta R}{R_0}\right) = \tfrac{2}{3}(\pi_{11}+2\pi_{12}-\pi_{44})\sigma_r \simeq -\tfrac{2}{3}\pi_{44}\sigma_r$$

The minor effect the stress system has on C is the favorable sense.

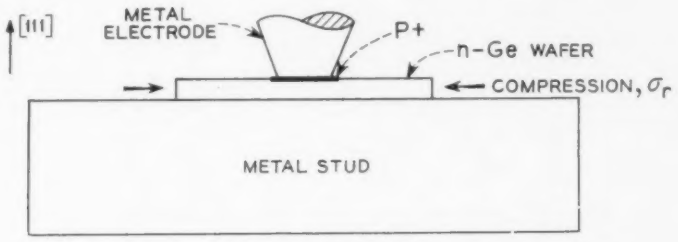


FIG. 2. Germanium varactor diode under compressive stress parallel to plane of junction.

6. OTHER APPLICATIONS

Effects of directional stress on the peak current of a silicon tunnel diode were reported[7] without elaboration, and effects of hydrostatic pressure on germanium diodes have also been described[8]. The effects are large enough to be of practical interest. However, a detailed analysis will not be attempted here.

One might increase the efficiency of a solar diode by imposing a stress that would increase the mobility of one or the other minority carrier in the direction normal to the p–n junction.

The principles illustrated above can be applied to other kinds of semiconductive devices, and in other ways to transistors or diodes. For example, the temperature sensitivity of some parameter of a device, such as the breakdown voltage of a Zener diode, might be reduced (or enhanced) by mounting the semiconductor in a way such that elastic stresses, proportional to the ambient temperature change, are produced by differing thermal expansivities of the semiconductor and its housing.

Speaking generally, if some parameter S is a linear function of temperature T and stress σ, it is in principle possible to compensate for changes in T if σ can be made to vary with T. Let:

$$A \equiv \left|\frac{\delta S}{\delta \sigma}\right|_T, \quad B \equiv \left|\frac{\delta S}{\delta T}\right|_\sigma \tag{10}$$

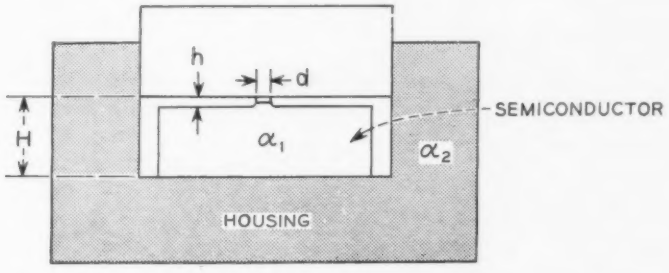


FIG. 3. Mesa diode housing for producing temperature-sensitive stress normal to plane of junction by differential thermal expansion.

Then, we require that

$$dS = A\,d\sigma + B\,dT = 0 \qquad (11)$$

From which the required temperature coefficient of stress is

$$\frac{d\sigma}{dT} = -\frac{B}{A} \qquad (12)$$

An example of a stress-magnifying housing, which utilizes the mesa form of semiconductor, is shown in Fig. 3. For such a housing, $d\sigma/dT$ is given approximately by

$$\frac{d\sigma}{dT} = \frac{{}^{1}E_1(\alpha_2-\alpha_1)H}{h+1{\cdot}8d} \qquad (13)$$

where α_2 and α_1 denote thermal expansivity of the housing wall and semiconductor, respectively, E_1 denotes Young's modulus of semiconductor, and h and d denote height and diameter of the mesa.

Table 2. Combinations of stress systems for silicon n–p–n and p–n–p transistors

Transistor type	Directions 3′	2′	1′	Stress system	$\Delta\mu_n(t)$	$\Delta\mu_p(l)$	$\Delta\mu_n(l)$	$\Delta\mu_p(t)$	ΔEg
1. *n–p–n*	[001]	[1$\bar{1}$0]	[110]	Compression in 2′	+ Good	+ Good		0 Good	− Bad
2. *n–p–n*	[001]	[1$\bar{1}$0]	[110]	Compression in 1′ and 2′	++ Good	0		0 Good	= Bad
3. *n–p–n*	[001]	[1$\bar{1}$0]	[110]	Tension in 1′ and 2′ (or compress. in 3′)	−− Bad	0		0	++ Good
4. *n–p–n*	[1$\bar{1}$0]	[001]	[110]	Compression in 1′	− Bad	+ Good		− Good	− Bad
5. *n–p–n*	[1$\bar{1}$0]	[001]	[110]	Compression in 1′ and 2′	+ Good	+ Good		− Good	− Bad
6. *n–p–n*	[111]	[1$\bar{1}$0]	[11$\bar{2}$]	Compression in 2′	∼0	+ Good		− Good	− Bad
7. *n–p–n*	[111]	[1$\bar{1}$0]	[11$\bar{2}$]	Compression in 1′ and 2′	∼0	+ Good		− Good	= Bad
8. *n–p–n*	[1$\bar{1}$0]	[111]	[11$\bar{2}$]	Compression in 1′	+ Good	+ Good		− Good	− Bad
9. *n–p–n*	[1$\bar{1}$0]	[111]	[11$\bar{2}$]	Compression in 1′ and 2′	+ Good	+(Small) Good		− Good	= Bad
10. *n–p–n*	[1$\bar{1}$0]	[111]	[11$\bar{2}$]	Tension in 2′	0	+ Good		− Good	+ Good
11. *p–n–p*	[111]	[1$\bar{1}$0]	[11$\bar{2}$]	Tension in 1′	0 Good		+ Good	+ Good	+ Good
12. *p–n–p*	[111]	[1$\bar{1}$0]	[11$\bar{2}$]	Tension in 1′ and 2′	∼0 Good		0 Good	++ Good	++ Good
13. *p–n–p*	[111]	[1$\bar{1}$0]	[11$\bar{2}$]	Compression in 3′	∼0 Good		∼0	++ Good	− Bad

Equation (13) assumes comparable magnitudes of Young's modulus and Poisson's ratio for semiconductor and housing wall. The factor, $H/(h+1{\cdot}8d)$ is a mechanical stress-amplification factor,[9] which increases the stress caused by differential expansion.

It is tempting to consider improving the performance of a semiconducting thermoelectric refrigerator or generator by using an elastically stressed single crystal, or polycrystals. The figure of merit commonly used is $Z \equiv Q^2/\rho k$, where Q is thermoelectric power, ρ is resistivity, k is thermal conductivity. By elastic stress we can decrease ρ, probably without a comparable increase in k. But what happens to Q? According to the many-valley model, decreasing the number, N_v, of valleys will decrease Q, as can be seen from HERRING's[10] equation for Q

$$Q = \mp 86{\cdot}2\left[\ln\frac{4{\cdot}70\times10^{15}}{n} + \tfrac{3}{2}\ln N_v + \tfrac{1}{2}\ln\left(\frac{m_1^*}{m}\cdot\frac{m_2^*}{m}\cdot\frac{m_3^*}{m}\right) + \tfrac{3}{2}\ln T + \frac{|\Delta\epsilon_T|}{RT}\right] \quad (\mu V/\text{deg}) \tag{14}$$

At optimum carrier concentrations n of order 10^{19}, the term $3\ln N_v/2$ is fairly dominant. Since stress would increase the mobility by emptying slow valleys, there is an effective reduction in N_v that tends to offset the increase in mobility. Nevertheless, the idea of operating a thermoelectric device with large elastic stresses in the semiconductor would appear to merit investigation.

Acknowledgements—The author is grateful to J. M. EARLY, H. L. FRISCH, C. HERRING, I. M. ROSS and R. WOLFE for helpful discussions and information.

REFERENCES

1. W. G. PFANN and R. N. THURSTON, *J. Appl. Phys.* **32**, 2008 (1961).
2. W. P. MASON and R. N. THURSTON, *J. Acoust. Soc. Amer.* **29**, 1096 (1957).
3. W. P. MASON, J. J. FORST and L. M. TORNILLO, *J. Instrum. Soc. Amer.* Preprint 15-NY60. Fall Meeting, New York (1960).
4. E. D. PADGETT and W. V. WRIGHT, *J. Instrum. Soc. Amer.* Preprint NY60-42. Fall Meeting, New York (1960).
5. W. PAUL, *J. Phys. Chem. Solids* **8**, 196 (1959).
6. L. J. GIACOLETTO, *RCA Rev.* **16**, 34 (1955).
7. L. ESAKI and Y. MIYAHARA, *Solid-State Electron.* **1**, 13 (1960).
8. S. L. MILLER, M. I. NATHAN and A. C. SMITH, *Phys. Rev. Letters* **4**, 60 (1960).
9. G. F. WEISSMANN. Private Communication (1960).
10. C. HERRING, *Bell Syst. Tech. J.* **34**, 237 (1955).

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 268–277. Printed in Great Britain

THE PERFORMANCE OF A CONVECTIVELY COOLED THERMOELEMENT USED FOR POWER GENERATION

CHARLES N. ROLLINGER* and J. EDWARD SUNDERLAND

Technological Institute, Northwestern University, Evanston, Illinois

(*Received* 12 *May* 1961; *in revised form* 14 *June* 1961)

Abstract—The differential equations which describe the steady-state temperature distribution in a thermoelement with constant properties and with convective heat transfer from its longitudinal surface are presented and solved by analytical methods. The closed-form analytical solutions are applied to a single-element thermoelectric generator to determine the effect of surface heat transfer upon the power output and efficiency over a range of values for the heat-transfer parameters.

The results show that in all cases the surface heat transfer increased the power output under matched load conditions; in many cases both the power output and the thermal efficiency were increased through the use of surface heat transfer. The conclusion is drawn that surface heat transfer can be beneficial in many applications, especially in cases where maximum power output is of primary importance, or in cases where the cold-junction fin area is greatly restricted by weight, space, or cost considerations.

Résumé—Les équations différentielles qui decrivent la distribution de température à l'état régulier dans un élément thermoélectrique ayant un transfert convectif de chaleur de sa surface longitudinale sont présentées et résolues par des méthodes analytiques. Les solutions analytiques à bornes fermées sont appliquées à un générateur thermoélectrique d'un élément pour déterminer l'effet du transfert de chaleur de surface sur la puissance de sortie et le rendement pour une gamme de valeurs affectées aux paramètres de transfert de chaleur.

Les résultats démontrent que dans tous les cas, le transfert de chaleur de surface a augmenté la puissance de sortie dans les conditions de charge equilibrée. Dans plusieurs cas, et le rendement thermoélectrique et la puissance de sortie ont été augmentés en utilisant le transfert de chaleur de surface. On conclut que le transfert de chaleur de surface peut être très utile dans plusieurs applications, surtout quand la puissance de sortie maximum est de première importance, ou dans les cas où la surface de bavure de la jonction froide est fortement restrainte par le poids, l'espace et d'autres considérations d'order économique.

Zusammenfassung—Die Differentialgleichungen, die die Temperaturverteilung im stationären Zustand für ein Thermoelement konstanter Eigenschaften und mit konvektiver Wärmeübertragung von seiner Oberfläche in Längerichtung beschreiben, werden aufgestellt und gelöst. Die analytischen Lösungen werden auf einen thermoelektrischen Generator mit einem Einzelelement angewendet, um den Einfluss des Wärmeverlustes von der Oberfläche auf die Energieausbeute und Leistung innerhalb eines gewissen Wertebereiches der Wärmeübertragungs-Parameter zu bestimmen.

Es ergibt sich in allen Fällen, dass der Wärmeverlust von der Oberfläche die Energieausbeute, bei Anpassung der Last (Widerstand des Thermoelements = Lastwiderstand) erhöhte; in vielen Fällen wurden sowohl Energieausbeute als auch die Wärmeleistung durch die Verwendung der Wärmeübertragung von der Oberfläche erhöht. Man kommt zu der Schlussfolgerung, dass Wärmeverlust von der Oberfläche in vielen Anwendungen von Nutzen sein kann, insbesondere in den Fällen, wo eine maximale Energieausbeute von grösster Wichtigkeit ist, oder in Fällen, wo für die Oberfläche der Rippen an der kalten Lötstelle in Bezug auf Gewicht, Raumverhältnisse oder Kosten Beschränkungen bestehen.

INTRODUCTION

THE use of semiconductor thermoelectric devices for the direct conversion of thermal energy to electrical energy has recently become practical in many specialized applications, but the relatively high first cost and low thermal efficiency of these

*Present address: Dept. of Mathematics, United States Air Force Academy, Colorado.

devices makes them as yet impractical for many applications. Improvements in performance can be gained as better materials are developed and as present design practices are improved, so that the maximum potential can be realized from available materials. This analysis is an attempt to determine if a better design can be achieved through the use of direct heat transfer between the thermoelement and its environment.

Surface heat transfer, i.e. heat transfer from the longitudinal surface of a thermoelement, has generally been considered undesirable under any condition of operation, but its actual effect upon the performance of a thermoelectric device has never been fully investigated. Previous analyses of thermoelectric devices which have been concerned primarily with performance optimization have assumed that the longitudinal surface is adiabatic, and, therefore, that all energy transfers between the thermoelement and its surroundings take place across the radial surfaces at the ends of the element. For example, the analyses of ALTENKIRCH[1], GELHOFF *et al.*[2], IOFFE[3] and BOERDIJK[4] have all assumed that the longitudinal surface of the element is adiabatic.

Analyses by HARMAN *et al.*[5], and KAGANOV *et al.*[6] have included surface heat transfer from a thermoelement to its environment. However, these analyses are concerned primarily with the determination of a theoretical basis for the accurate measurement of the thermoelectric properties of the thermoelement. The results of these analyses do not indicate what effect the surface heat transfer has upon the performance of a useful thermoelectric device.

A recent analysis by BRANDT[7] did consider the effect of radiation heat transfer from the longitudinal surface upon the performance of a thermoelement used for power generation. The differential equation for the model which Brandt studied was nonlinear and an analytical solution was not determined. However, by numerical analysis he determined the effect of the surface heat transfer upon the thermal efficiency for one set of conditions and found for this case that the radiation heat transfer significantly lowered the thermal efficiency.

The object of the present analysis is to determine the effect of surface heat transfer upon the performance of a thermoelement used as a power generator.* For the model studied, analytical expressions can be obtained for the power output and the thermal efficiency. These analytic expressions are useful for studying the effects of surface heat transfer upon the performance of a thermoelement.

THE PHYSICAL SYSTEM AND ITS MATHEMATICAL MODEL

The physical system considered in this analysis is a single-element thermoelectric power generator with a fixed hot-junction temperature T_H and a fixed environment temperature T_E. The system is shown schematically in Fig. 1. Surface heat transfer is assumed to take place by convection, but radiation heat transfer could also be treated in this linear form with the methods discussed by CARSLAW and JAEGER[8]. The surface of the arm between $x = 0$ and $x = a$ is assumed to be perfectly insulated; the surface of the arm between $x = a$ and $x = L$ transfers heat by convection so that the heat loss dQ is given by

$$dQ = hp(T - T_E)dx \tag{1}$$

where p is the perimeter of the thermoelement and T is the temperature at any position x. The heat-transfer coefficient h is constant in the region $a \leqslant x \leqslant L$. The surface area of the element exposed to convection varies from zero, when $a = L$, to the entire surface area of the element when $a = 0$. In addition it is assumed that the cross-sectional area A, the Seebeck coefficient S, the (isothermal) electrical resistivity ρ, and the thermal conductivity (for zero electrical current) K of the thermoelement are constant. It is also assumed that the electrical leads which complete the circuit have zero thermal conductivity, Seebeck coefficient and electrical resistivity. If the temperature variations in the radial direction are assumed to be small compared with variations in the axial direction, the differential equations for the steady-state temperature distribution in the thermoelement are

$$KA\frac{d^2T_1}{dx^2} + I^2\rho/A = 0, \quad 0 \leqslant x \leqslant a \tag{2}$$

* The effect of surface heat transfer on the performance of a thermoelement used as a Peltier heat pump is currently under investigation.

and

$$KA\frac{d^2T_2}{dx^2} - hp(T_2 - T_E) + I^2\rho/A = 0, \quad a \leqslant x \leqslant L \tag{3}$$

In equation (3), the first term results from heat conduction in the element; the second term, made at the cold junction in order to determine T_C. Neglecting the effects of thermal and electrical contact resistances, an energy balance at the cold junction results in the expression

$$-KA\frac{dT_2}{dx}\bigg|_{x=L} + IST_C = h_f A_f (T_C - T_E) \tag{5}$$

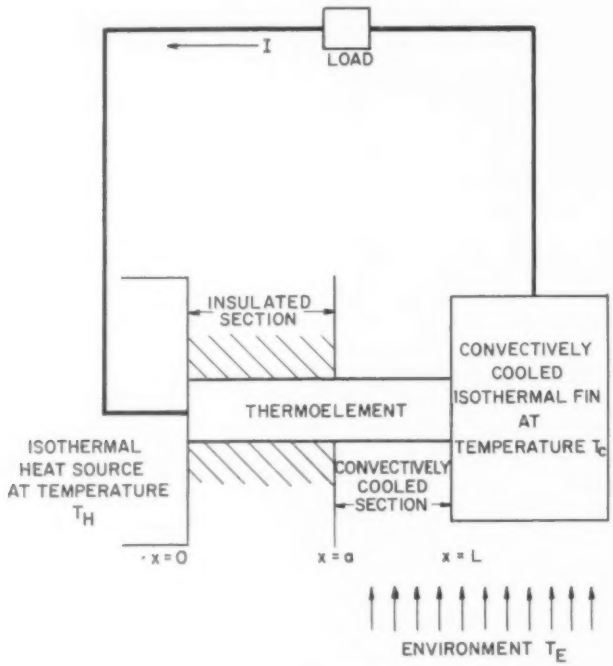


FIG. 1. Single-element thermoelectric generator.

convective heat transfer from the element; and the third term, Joulean heat produced in the element. These two differential equations are subject to the following boundary conditions

$$T_1(0) = T_H \tag{4a}$$

$$T_1(a) = T_2(a) = T_a \tag{4b}$$

$$T_2(L) = T_C \tag{4c}$$

$$\frac{dT_1}{dx}\bigg|_{x=a} = \frac{dT_2}{dx}\bigg|_{x=a} \tag{4d}$$

Since the cold-junction temperature T_C is not prescribed directly, an energy balance must be

where the entire surface area of the fin, A_f, is assumed to remain at the temperature T_C. The first term is the heat conduction from the element, the second is the Peltier heat liberated at the cold junction, and the sum of these must be equal to the heat transferred by convection from the fin to the environment. Note that the fin heat-transfer coefficient h_f need not have the same value as the element heat-transfer coefficient h.

Equations (2) and (3) with boundary conditions (4) and (5) are sufficient to solve for the two unknown temperature distributions, T_1 and T_2. In order to reduce the number of parameters, equations (2) to (5) will be put into dimensionless

form. Let the dimensionless temperature θ be defined as

$$\theta = (T-T_E)/(T_H-T_E) \tag{6}$$

the dimensionless position variable u be defined as

$$u = \frac{x}{L}\left[\frac{hpL^2}{KA}\right]^{1/2} \tag{7}$$

and the dimensionless current Y be defined as

$$Y = ISL/KAu_L \tag{8}$$

where u_L is u evaluated at $x = L$. Substitution of these dimensionless quantities into equations (2) to (5) transforms them to

$$\frac{d^2\theta_1}{du^2} + Y^2\alpha = 0,\ 0 \leqslant u \leqslant u_a \tag{9}$$

$$\frac{d^2\theta_2}{du^2} - \theta_2 + Y^2\alpha = 0,\ u_a \leqslant u \leqslant u_L \tag{10}$$

and where Z is the figure of merit of the material. The solution of equation (9) subject to boundary conditions (11a, b) is

$$\theta_1(u) = 1 + \left[\frac{\theta_a - 1}{u_a} + \tfrac{1}{2}u_a Y^2\alpha\right]u - \tfrac{1}{2}Y^2\alpha u^2 \tag{14}$$

The solution of equation (10) subject to boundary conditions (11b, c) is

$$\theta_2(u) = Y^2\alpha + [\theta_C - Y^2\alpha]\exp(u-u_L) + \left[\frac{\theta_a - \theta_C e^U - Y^2\alpha(1-e^U)}{\sinh U}\right]\sinh(u-u_L) \tag{15}$$

where

$$U = u_a - u_L = u_L[(a/L)-1] \tag{16}$$

Applying boundary condition (11d) to equations (14) and (15) and solving for the temperature at $x = a$,

$$\theta_a = \frac{u_a\theta_C e^U[1-\tanh U] - \tanh U - [Y^2\alpha u_a \tanh U][\tfrac{1}{2}u_a - \tanh U/2]}{u_a - \tanh U} \tag{17}$$

Solving equations (12), (15) and (17) for the temperature at the cold junction,

$$\theta_C = \frac{Y^2\alpha[u_a(\tfrac{1}{2}u_a - \tanh U/2)\operatorname{sech} U - (u_a - \tanh U)\tanh U/2] + Y\phi(u_a - \tanh U) + \operatorname{sech} U}{(1 - Y + H_f/u_L)(u_a - \tanh U) + e^U(1-u_a)\operatorname{sech} U} \tag{18}$$

$$\theta_1(0) = 1 \tag{11a}$$

$$\theta_1(u_a) = \theta_a = \theta_2(u_a) \tag{11b}$$

$$\theta_2(u_L) = \theta_C \tag{11c}$$

$$\left.\frac{d\theta_1}{du}\right|_{u=u_a} = \left.\frac{d\theta_2}{du}\right|_{u=u_a} \tag{11d}$$

$$-u_L\left.\frac{d\theta_2}{du}\right|_{u=u_L} + u_L Y(\theta_C + \phi) = H_f\theta_C \tag{12}$$

where α, ϕ and H_f are defined as

$$\alpha = \frac{\rho K}{S^2(T_H - T_E)} = \frac{1}{Z(T_H - T_E)} \tag{13a}$$

$$\phi = T_E/(T_H - T_E) \tag{13b}$$

$$H_f = h_f A_f L/KA \tag{13c}$$

Equations (14), (15), (17) and (18) can be used to compute the temperature distribution in the element for any values of the dimensionless parameters α, ϕ, u_L, H_f, u_a and Y.

DETERMINATION OF PERFORMANCE CRITERIA

There are two performance criteria which are of importance for a power generator. These are the power delivered to load and the thermal efficiency. The power output for a generator can be expressed as

$$P = I^2R \tag{19}$$

where R is the resistance of the load. In this analysis it is assumed that the load resistance is equal to the resistance of the thermoelement which gives the expression

$$P = I^2\rho L/A \tag{20}$$

This condition, which will be referred to as matched load condition, will give the maximum power output only for the case when H_f is infinite, i.e. when the cold-junction temperature is independent of the current. If H_f is finite the condition of maximum power output does not occur when the load resistance is equal to the resistance of the thermoelement. The determination of a general expression for the true maximum power output presents some analytical difficulties, but numerical calculations have shown that the difference between the true maximum power output and the power output under matched load conditions is small. For the values of parameters used in this analysis the difference is about 0·1 per cent of $P_{\max}$ at H_f equal to 13, 0·3 per cent of $P_{\max}$ for H_f equal to 5, and 7 per cent of $P_{\max}$ for H_f equal to $\frac{2}{10}$. The thermal efficiency is defined as the ratio of useful work to the energy leaving the heat source. Thus,

$$\epsilon = \frac{I^2\rho L/A}{-KA\left.\frac{dT}{dx}\right|_{x=0} + IST_H} \tag{21}$$

The numerator is the electrical power consumed within the device which under matched load conditions is equal to the power output. The first term in the denominator is the heat removed from the hot junction due to conduction through the thermoelement, and the second term is the amount of heat removed from the hot junction due to the Peltier effect.

Expressing equations (20) and (21) in non-dimensional form,

$$P^* = Y^2\alpha u_L^2 \tag{22}$$

and

$$\epsilon = \frac{Y^2\alpha u_L^2}{-u_L\left.\frac{d\theta}{du}\right|_{u=0} + Yu_L(\phi+1)} \tag{23}$$

where P^* is the dimensionless power output defined by the relation

$$P^* = \frac{PL}{KA(T_H - T_E)}$$

Note that the temperature gradient in the denominator of equation (23) is evaluated from equation (15) if $u_a = 0$ and from equation (14) if $u_a > 0$.

If the load resistance is equal to the resistance of the thermoelement, then

$$\tfrac{1}{2}S(T_H - T_C) = I\rho L/A \tag{24}$$

which in dimensionless form becomes

$$\tfrac{1}{2}(1-\theta_C) = u_L Y\alpha \tag{25}$$

Solving equations (18) and (25), the current Y under matched load conditions is

$$Y = \frac{-F-[F^2-4EG]^{1/2}}{2E} \tag{26a}$$

where

$$E = \alpha[(u_a - \tanh U)(2u_L + \tanh U/2) - u_a(\tfrac{1}{2}u_a - \tanh U/2)\operatorname{sech} U] \tag{26b}$$

$$F = -(u_a - \tanh U)[\phi + 2\alpha(H_f + u_L) + 1] + 2\alpha u_L e^U(u_a - 1)\operatorname{sech} U \tag{26c}$$

$$G = (u_a - \tanh U)(1 + H_f/u_L) - [1 + e^U(u_a - 1)]\operatorname{sech} U \tag{26d}$$

The negative value of the square root is used because it yields the solution for a practical power generator. The positive root yields a very large value of current which causes the absolute temperature of the cold junction to be negative. It is obvious that this condition is physically impossible.

The smallest possible value of θ_C is zero. This condition exists when $T_C = T_E$. Under this condition of operation the current Y will be a maximum and will have a value

$$Y_{\max} = 0{\cdot}5/\alpha u_L \tag{27}$$

The corresponding maximum power output will be

$$P^*_{\max} = 0{\cdot}25/\alpha \tag{28}$$

The maximum thermal efficiency $\epsilon_{\max}$ for matched load occurs when the element is completely insulated ($a/L = 1$) and when $\theta_C = 0$. Under these conditions

$$\epsilon_{\max} = \frac{1}{4\alpha - 0{\cdot}5 + 2(\phi+1)} \tag{29}$$

RESULTS

Equations (14) to (18), (22), (23), (25) and (26) can be used to determine the temperature distribution in the element $\theta(u)$, the power output P^* and the thermal efficiency ϵ, under varying conditions of surface heat transfer. Because of the complexity of the expressions for P^* and ϵ, it is not possible to determine by simple comparisons what effect a change in surface heat transfer has upon the performance of the generator.

In order to determine the effect of surface heat transfer on P^* and ϵ, specific values will be substituted for the parameters. The values of the physical constants used to calculate these parameters are taken from BRANDT[7]; these values are

$$S = 200\times10^{-6}\ V/°K \qquad p = 1{\cdot}92\ \text{cm}$$
$$\rho = 3\times10^{-3}\ \Omega\text{-cm} \qquad L = 2{\cdot}54\ \text{cm}$$
$$K = 14\times10^{-3}\ \text{W/cm°K} \qquad T_H = 812°\ K$$
$$A = 0{\cdot}292\ \text{cm}^2 \qquad T_E = 300°\ \text{K}$$

These values resulted in the following values for α and ϕ: $\alpha = 2{\cdot}051$ and $\phi = 0{\cdot}586$.

Two values were used for the convective heat-transfer coefficient h. These are $h = 11{\cdot}3\times10^{-4}$ W/cm^2 ° K which is the value used by BRANDT and which corresponds to a heat-transfer coefficient for natural convection, and $h = 56{\cdot}8\times10^{-4}$ W/cm^2 ° K which corresponds to a value of heat-transfer coefficient for forced convection with air. These two values of h, together with the properties of the material and the geometric dimensions, yield

$$u_L = 1{\cdot}85 \quad \text{for } h = 11{\cdot}3\times10^{-4}$$
$$u_L = 4{\cdot}148 \quad \text{for } h = 56{\cdot}8\times10^{-4}$$

The quantity a/L varied from 0, when the surface of the arm is completely insulated, to 1, when heat transfer takes place over the entire surface of the arm. The range of H_f was 0–15. The value 0 indicates that the radial surface at the end of the element is insulated; the value of 15 corresponds to the condition of very low thermal resistance across the fin. It is important to note that all calculations were made assuming that the entire fin surface is at the temperature T_C which occurs only if the thermal conductivity of the fin is extremely large.

Under these conditions, $\theta(u)$, P^* and ϵ were computed from the analytical expressions which have been presented. The results of these computations are shown graphically in Figs. 2–7.

DISCUSSION OF THE RESULTS

Figure 2 shows how convection along the surface affects the temperature distribution under matched load conditions and for one particular set of values for the other parameters.

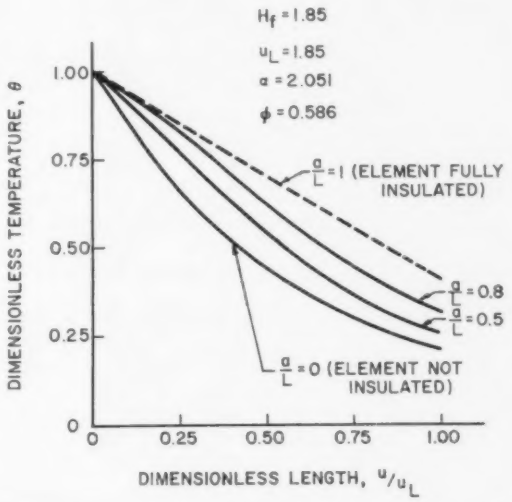


FIG. 2. Temperature distribution for matched load conditions.

Figure 3 indicates the effect of surface heat transfer on the power output of the generator over a range of values for H_f. The quantity H_f can be physically interpreted as the heat conductance of the fin; H_f is proportional to the surface area of the fin. This plot indicates that, for any given value of H_f, the power output is always increased as the surface area of the element exposed to the environment is increased. From Fig. 2 it can be seen that convection along the surface of the element causes T_C to decrease. Since the voltage generated by the Seebeck effect is proportional to $(T_H - T_C)$, this will result in an increase in power output of the generator.

It is interesting to note from Fig. 3 that removing one half the insulation ($a/L = 0{\cdot}5$) in most cases results in an increase in power output equal to

about 80 per cent of the total increase in power output that could be gained by removing all the insulation. The real significance of this fact can be seen in Fig. 4 which is a plot of the thermal efficiency vs. H_f.

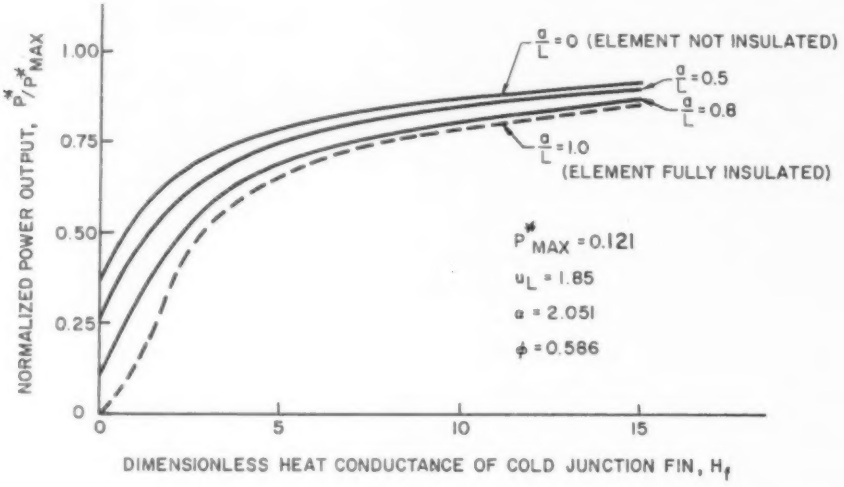


FIG. 3. Power output vs. heat conductance of the cold-junction fin for various values of a/L.

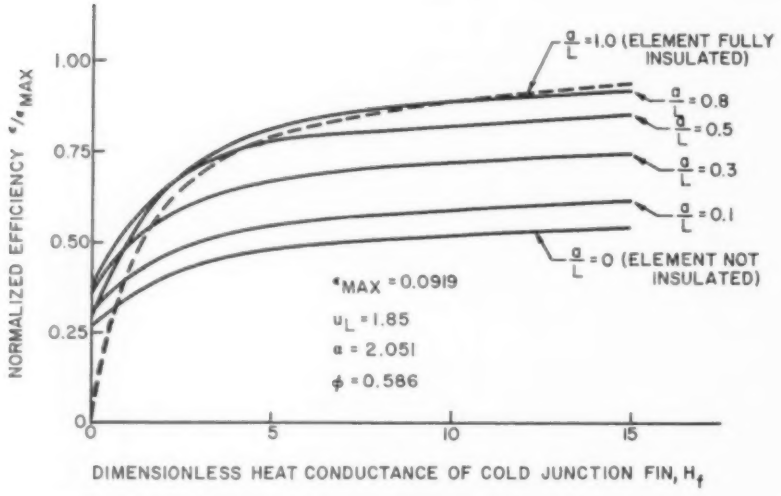


FIG. 4. Efficiency vs. heat conductance of the cold-junction fin for various values of a/L.

Figure 4 shows that the thermal efficiency may either increase or decrease as a/L decreases. Note that in this plot the $a/L = 0{\cdot}5$ line lies close to the $a/L = 1{\cdot}0$ line. This means that, by removing half the insulation from the arm, about 80 per cent of the maximum gain in power output can be achieved with only a comparatively slight loss of thermal efficiency. Fig. 4 also shows that if $H_f = 4{\cdot}25$, the efficiencies for $a/L = 1{\cdot}0$ and

$a/L = 0{\cdot}5$ are equal. At the same value of H_f, Fig. 3 shows that the element with $a/L = 0{\cdot}5$ is producing about 30 per cent more power than the element with $a/L = 1{\cdot}0$. Thus, at this value of H_f, removing insulation from half of the element would increase the power output by 30 per cent without reducing the thermal efficiency.

It can also be seen from Fig. 4 that H_f must be quite large before a completely insulated element will have a greater efficiency than one that is cooled along its surface. In the case shown H_f must be greater than 12 before the efficiency of the completely insulated element exceeds the efficiency of the convectively cooled elements. For H_f lower than 12, more power output and an increase in thermal efficiency can be gained by removing some portion of the insulation from the element. The exact value of a/L which will give the highest efficiency under matched load conditions for any value of the parameters can be computed from equation (23).

Figure 5 is a plot of thermal efficiency for matched load conditions vs. a/L for various values of H_f. It can be seen that the value of a/L which gives the maximum efficiency for $H_f = 0$ is near $a/L = 0{\cdot}5$ and that as H_f increases this value moves toward $a/L = 1{\cdot}0$. Thus, for large values of H_f, the maximum efficiency occurs if the element is completely insulated. Regardless of the value of H_f, the maximum power output occurs if the element is convectively cooled over its entire surface.

Figures 6 and 7 show the effect of u_L on the power output and thermal efficiency respectively. It can be seen that the performance of the insulated

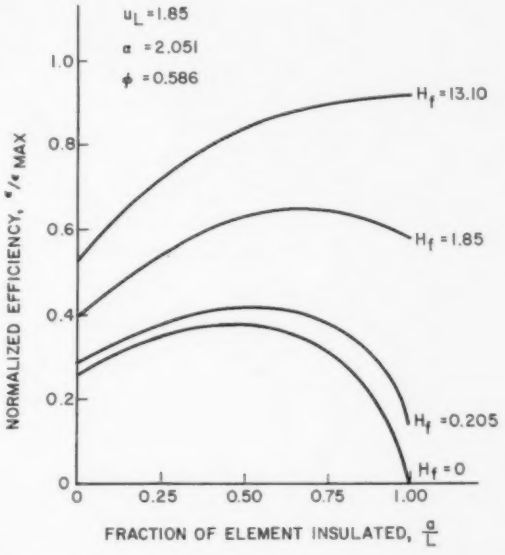


FIG. 5. Efficiency vs. fraction of thermoelement insulated

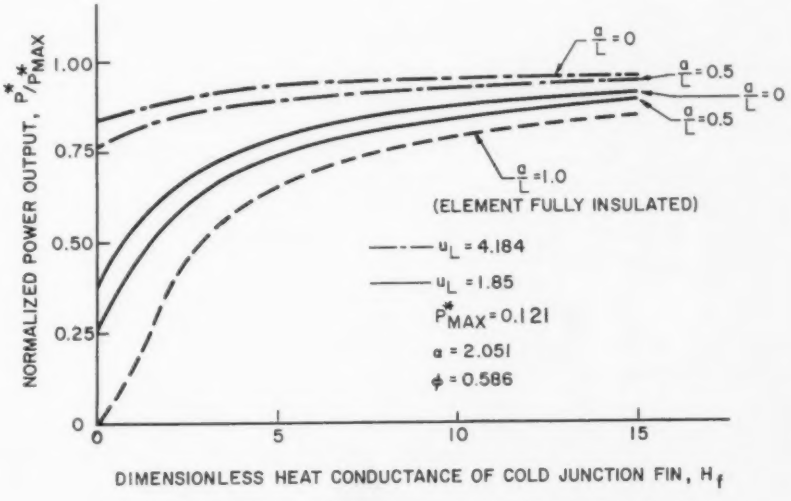


FIG. 6. Power output vs. heat conductance of the cold-junction fin for various values of u_L and a/L.

thermoelement does not depend on u_L. Fig. 6 indicates that raising the value of u_L makes surface heat transfer more advantageous from the power-output viewpoint; Fig. 7 shows that a higher value of u_L degradates the thermal efficiency for a thermoelement with surface heat transfer.

reduction in fin area or the number of elements could result in a considerable savings in the over-all weight and size of the device.

Few designs are based completely on either maximum power output or on thermal efficiency, but rather on some combination of these factors.

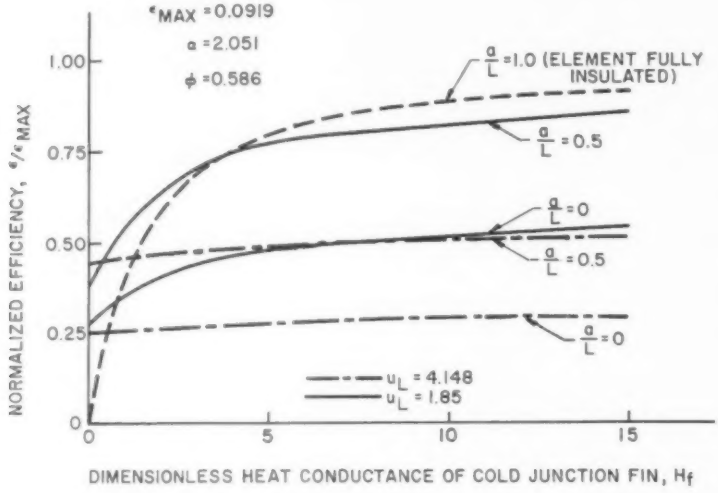


FIG. 7. Efficiency vs. heat conductance of the cold-junction fin for various values of u_L and a/L.

For any other values of α and ϕ, curves such as those shown on Figs. 6 and 7 can be constructed by using equations (22), (23) and (26). These curves can then be used to indicate how much fin surface area and surface heat transfer should be used for any given design.

CONCLUSIONS

The results that have been presented indicate that surface heat transfer appears to be a useful mechanism for increasing the power output of a thermoelectric power generator under any condition of operation. The results further indicate that in many cases both the power output and the thermal efficiency can be increased by the use of this mechanism.

By using convective heat transfer along the surface of the thermoelements in a multi-element generator, a given power output could be achieved with less total fin area at the cold junction or by using fewer elements in the generator. This

The analytic expressions presented can be used to determine the amount of surface heat transfer that is most beneficial for a particular application. The conclusion can be drawn that surface heat transfer will be most beneficial in applications where maximum power output is of primary importance, or in applications where there are severe restrictions upon the allowable cold-junction fin system resulting from weight, space or cost considerations.

Acknowledgement—The authors gratefully acknowledge the help of Mr. R. W. SPRINGER who performed many of the calculations.

REFERENCES

1. E. ALTENKIRCH, *Phys. Z.* **12**, 920 (1911).
2. P. O. GELHOFF, E. JUSTI and M. KOHLER, *Abh. braunschw. wiss. Ges.* **2**, 149 (1950).
3. A. F. IOFFE, *Semiconductor Thermoelements and Thermoelectric Cooling*. Infosearch, London (1957).

4. A. H. Boerdijk, *J. Appl. Phys.* **30,** 1080 (1959).
5. T. C. Harman, J. H. Cahn and M. J. Logan, *J. Appl. Phys.* **30,** 1351 (1959).
6. M. A. Kaganov, I. S. Lisker and I. B. Mushkin, *Soviet Phys. Solid State* **1,** 905 (1959).
7. J. A. Brandt, *Solutions to the Differential Equations Describing the Temperature Distribution, Thermal Efficiency, and Power Output of a Thermoelectric Element with Variable Properties and Cross-Sectional Area.* Paper presented at the Joint Technical Society, Department of Defense Symposium on Thermoelectric Energy Conversion, Dallas, Texas, January 9, 1961.
8. H. S. Carslaw and J. E. Jaeger, *Conduction of Heat in Solids,* (2nd ed.) p. 21. Oxford University Press, London (1959).

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 278–290. Printed in Great Britain

A SIMPLIFIED APPROACH TO TRANSISTOR ADMITTANCES

J. LINDMAYER and C. WRIGLEY

Research Laboratories, Sprague Electric Co., North Adams, Massachusetts

(*Received* 14 *June* 1961)

Abstract—Simple admittance parameters, which are valid for a limited frequency range, can describe high-frequency transistors relatively well. This is possible since the actual cutoff frequency is much lower than the cutoff frequency arising from the base transport. Two charges play the dominant role in determining the admittances: the stored charge, and the portion of it reclaimable through the emitter during discharge. For a constant built-in field in the base region the treatment is so simple that the admittance parameters can be written down by inspection. It is shown that the ratio of the transfer admittances is a real number for any transistor and all frequencies. This constant is determined in general for an arbitrary impurity distribution. The diffusion capacitances are discussed as well as practical problems associated with their measurement.

Résumé—Des simples paramètres d'admittance qui sont valables pour une gamme d'admittance limitée peuvent relativement bien décrire les transistors à haute frequence. Ceci est possible grâce au fait que la fréquence de coupure actuelle est beaucoup plus inférieure à la fréquence de coupure due au transport de base. Deux charges jouent un rôle prépondérant pour déterminer les admittances; la charge emmagasinée, et une portion de celleci qui est récupérée à travers l'émetteur durant la décharge. Pour une constante de champ fixé dans la région de base, l'analyse est si simple que les paramètres d'admittance peuvent être obtenus par observation.

Il est démontré que le rapport des admittances de transfert est un nombre réel pour un transistor à toutes les fréquences. Cette constante est determinée en général par une distribution arbitraire d'impuretés.

On traite aussi des capacités de diffusion et des problèmes pratiques associés aux mesures de ces dernières.

Zusammenfassung—Hochfrequenz-Transistoren lassen sich durch einfache, in einem beschränkten Frequenzbereich gültige Scheinleitwert-Parameter recht gut beschreiben. Dies ist möglich, weil die tatsächliche Grenzfrequenz viel niedriger ist als die durch den Transport in der Basiszone entstehende Grenzfrequenz. Zwei Ladungen sind für die Bestimmung der Scheinleitwerte entscheidend: die gespeicherte Ladung und der Anteil dieser Ladung, der während der Entladung vom Emitter zurückgewonnen werden kann. Bei konstantem Feld in der Basiszone ist die Berechnung so einfach, dass die Scheinleitwert-Parameter unmittelbar gefunden werden können.

Es wird gezeigt, dass das Verhältnis der Übertragungsscheinleitwerte bei jedem Transistor und bei allen Frequenzen eine reelle Zahl ist. Diese Konstante wird allgemein für eine willkürliche Störstellenverteilung bestimmt.

Dann folgt eine Erörterung der Diffusions-Kapazitäten und der mit ihrer Messung verbundenen praktischen Probleme.

1. INTRODUCTION

THE admittance parameters describing junction-transistor behavior for high frequencies have been calculated by SHOCKLEY[1] for diffusion transistors and by KRÖMER[2] for drift transistors. The solutions lead to hyperbolic forms containing complex arguments and are difficult to handle. BEAUFOY and SPARKES[3] introduced an analysis by charge control, which applies for a limited frequency range, but greatly simplifies the treatment. The following argument based on currents and charges makes possible an extremely simple treatment, including the drift transistor.

In today's high-frequency transistors, a limited frequency range is sufficient to describe transistor behavior. To emphasize this point let us consider,

for example, that the actual α-cutoff frequency of a transistor is[4]

$$\frac{1}{f_{c\alpha}} = \frac{1}{f_{c\alpha}^*} + \frac{1}{f_e} + \frac{2}{f_c}$$

The frequency $f_{c\alpha}^*$ is the α-cutoff frequency of the "ideal" transistor, f_e is a cutoff frequency controlled by the emitter space-charge capacitance, and f_c is a cutoff frequency determined by the collector space-charge capacitance and the base resistance. These three frequencies result in the actual cutoff frequency of the transistor, $f_{c\alpha}$. For today's high-frequency transistors the "ideal" cutoff frequency is in the kilomegacycle range, while the actual cutoff usually occurs in the hundred megacycle range. This difference by an order of magnitude means that such transistors are operated well below the ideal cutoff. This, in turn, allows us to describe the ideal transistor by parameters which apply only for a limited range of frequencies.

The parameters of the ideal transistor will be calculated first, to which the space-charge capacitances and the base resistance can then be added. An exponential impurity distribution describes a diffused base region quite well, and its constant field results in an extremely simple treatment.

2. NOTATION

A junction area (cm²)
C_c collector space-charge capacitance (F)
C_D input diffusion capacitance in common emitter (F)
C_{DC} collector diffusion capacitance (F)
C_{DE} emitter diffusion capacitance (F)
C_E emitter space-charge capacitance (F)
C_m input capacitance measured in common emitter (F)
D diffusion constant of holes (cm²/sec)
D_n diffusion constant of electrons (cm²/sec)
E_b built-in field in the base (V/cm)
I_c collector current (A)
I_e emitter current (A)
I_E d.c. emitter current (A)
I_n electron current (A)
I_p hole current (A)
I_R emitter discharge current (holes) (A)
j $\sqrt{-1}$
k Boltzmann constant (V−C/°K)
K_c ratio of transfer admittance (dimensionless)
n electron concentration (cm⁻³)
N impurity concentration (cm⁻³)
N_E impurity concentration at the emitter (cm⁻³)
N_w impurity concentration at W (cm⁻³)
p hole concentration (cm⁻³)
p_e hole concentration at the emitter (cm⁻³)
q electronic charge (C)
Q_e charge reclaimable through the emitter (C)
Q_{er} same as Q_e with an opposite field (C)
Q_t charge involved in inductive effects (C)
Q_s charge stored in the base (C)
T temperature (°K)
V_c collector voltage (V)
V_e emitter voltage (V)
V_E emitter d.c. voltage (V)
W basewidth (cm)
x_0 $1/e$ length of the exponential (cm)
y_{11} input admittance (Ω⁻¹)
y_{21} forward transfer admittance (Ω⁻¹)
y_{12} reverse transfer admittance (Ω⁻¹)
y_{22} output admittance (Ω⁻¹)
α common-base current gain (dimensionless)
α_0 low-frequency common-base current gain (dimensionless)
β common-emitter current gain (dimensionless)
β_0 low-frequency current gain $= \alpha_0/(1-\alpha_0)$ (dimensionless)
η normalized built-in field (dimensionless)
μ mobility of holes (cm²/V-sec)
μ_n mobility of electrons (cm²/V-sec)
τ_B base transit time (sec)
ω angular frequency (sec⁻¹)

3. THE STORED CHARGE

The junction transistor is represented schematically in Fig. 1. The impurity distribution, $N(x)$, in the base region between the emitter and collector space-charge boundaries is taken to be exponential.

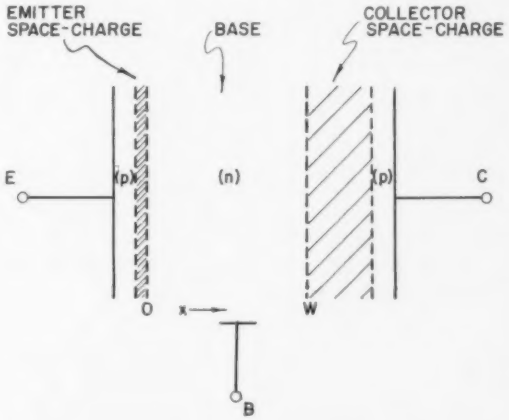


FIG. 1. Schematic representation of the junction transistor.

$$N(x) = N_E \exp(-x/x_0) = N_E \exp(-\eta x/W) \tag{1}$$

where N_E is the concentration of impurities at the emitter space-charge boundary and W is the base-width between space charge boundaries. The symbol η is introduced for convenience, and from (1) it is

$$\eta = \frac{W}{x_0} = \ln \frac{N_E}{N_w} \tag{2}$$

η is a measure of the built-in field in the base region, as is shown in Appendix I. Apparently $\eta = 0$ is the case of a diffusion transistor with no built-in field. For drift transistors η is in the range 3–8.

If we consider a p–n–p transistor for convenience, the holes injected through the emitter barrier represent a constant current across the base. Today's transistors have narrow base regions, and recombination in the base is therefore negligible. The current at every point is the sum of the diffusion and drift currents.

$$Ip = -qADp' + qAD\frac{\eta}{W}p \tag{3}$$

where the field expression in Appendix I was utilized, and the prime refers to differentiation with respect to x.

Solving (3) for p, with the boundary condition $p = 0$ at $x = W$, we obtain

$$p(x) = \frac{I_p W}{qAD} \frac{1 - \exp[-\eta(1-x/W)]}{\eta} \tag{4}$$

This distribution is plotted in Fig. 2.

The stored charge in the base region is obtained by integrating (4) from $x = 0$ to $x = W$.

$$Q_s = qA \int_0^W p\,dx = \frac{I_p W^2}{D} \frac{\eta - 1 + \exp(-\eta)}{\eta^2} \tag{5}$$

4. ADMITTANCES

The admittance parameters of the transistor are determined by the variation of the stored charge by the emitter and collector voltages. The four-pole admittance definitions are shown in Fig. 3 for reference.

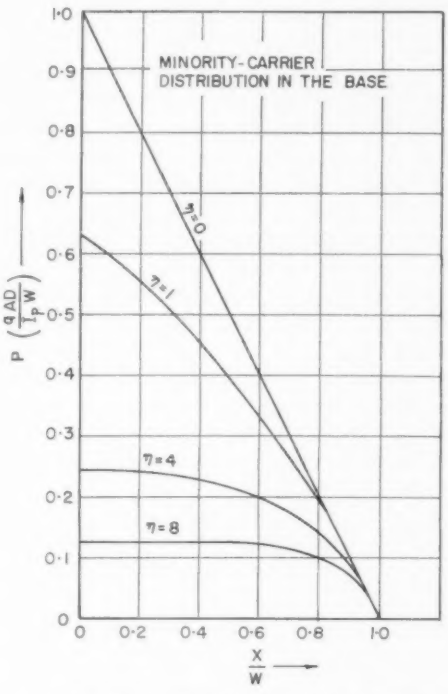
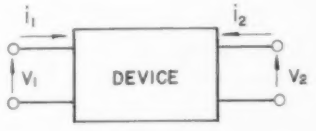


FIG. 2. Normalized hole distribution in the base of a p–n–p transistor. As the field is increased holes are swept away from the emitter.



$$i_1 = y_{11} V_1 + y_{12} V_2$$

$$i_2 = y_{21} V_1 + y_{22} V_2$$

FIG. 3. Four-pole (two-port) admittance parameter definitions.

A general relationship between two of the admittance parameters can be found promptly. If the collector were replaced by an ohmic contact there would be no change in the hole distribution,

(4), since the ohmic contact provides the same boundary condition.

It is apparent that, without the collector space-charge region, no gain in energy would be available. Because we would then have a passive device, the forward and reverse transfer admittances would be equal, $y_{21} = y_{12}$. This holds for all passive networks. The gain must arise because of the collector space-charge region. In the common base configuration it just happens that the "passive" base and "active" collector regions form a "chain", (Fig. 1). If the voltage gain due to the collector space-charge layer is denoted by K_c, we must have

$$\frac{y_{21B}}{y_{12B}} = K_c \tag{6}$$

which occurs when a voltage amplifier is in chain with a passive network. This *simple relationship applies to all transistors at all frequencies.*† K_c is calculated in Appendix IV.

There is another important facet of (6) which will be useful. Let us write (6) in the differential form resulting from the definitions of y_{12B} and y_{21B}.

$$\begin{aligned} K_C &= \frac{\partial I_c}{\partial V_e}\bigg|_{dV_c=0} \frac{\partial V_c}{\partial I_e}\bigg|_{dV_e=0} \\ &= \frac{\partial I_{pc}}{\partial V_e}\bigg|_{dV_c=0} \frac{\partial V_c}{\partial I_{pe}}\bigg|_{dV_e=0} \end{aligned} \tag{7}$$

We could replace I_c by I_{pc}, since the collector current is a hole current. On the other hand I_e could be replaced by I_{pe}, since the constant emitter voltage allows no change in the electron current.

Now, consider that K_c is real (since the transit time through the space charge is short) and that the currents in (7) are the complex dependent variables. In order that K_c be real it must happen that $I_{pe} = I_{pc}$ for all frequencies in (7). Then we have a general relationship between the short-circuit voltage operators

$$\frac{d}{dV_e} = K_c \frac{d}{dV_c} \tag{8}$$

Equipped with (6) and (7), we can easily calculate the admittances. Let us first determine the input admittance, y_{11B}. The emitter current is the sum of two currents (time invariant and time dependent). The former is the sum of the d.c. hole and electron currents, i.e. the d.c. emitter current.

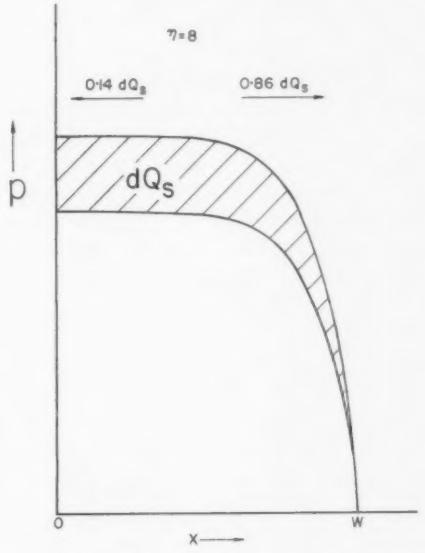


FIG. 4. The change of the hole distribution from one steady state to another for a $-dV_e$ change in the emitter voltage. The portions of the extra charge leaving in each direction are shown.

The time-dependent current is the time derivative of a charge. This charge, denoted Q_e, is the part of Q_s reclaimable through the emitter lead when the emitter voltage is dropped to zero. Q_e is derived in Appendix II. This is the only additional charge which need be calculated in order to obtain the admittances. Fig. 4 shows the variation in the stored charge for a change in the emitter voltage.

† While the above argument may seem rather superficial, its validity is supported by the following: (a) ZAWELS[5] has shown that a K_c amplifier can be used in the equivalent circuit of a diffusion transistor; (b) if KRÖMER's admittances[2] are examined it is found that the ratio in (6) is obtained as (IV.8) in Appendix IV; (c) it can be shown[6] in general that the minority-carrier flow in the base is analogous to a current flow in a distributed passive *RC* network; (d) if calculation of the admittances is made[6] by separating the passive base and the K_c amplifier for an exponential impurity distribution, KROEMER's[2] results are obtained.

From the above argument we can write the emitter current as

$$I_e = I_E + \frac{dQ_e}{dt} = I_E + j\omega Q_e \tag{9}$$

The current through a retarding barrier follows the relation

$$I_p = \alpha_0 I_E = A_0 \exp(qV_E/kT) \tag{10}$$

where α_0 is determined by the injection efficiency and surface recombinations, and A_0 is a proportionality constant.

The input admittance is now obtained from its definition, using (9) and (10),

$$y_{11B} = \left.\frac{\partial I_e}{\partial V_e}\right|_{dV_c=0} = \frac{q}{kT}(I_E + j\omega Q_e) \tag{11}$$

The other three admittances can be obtained easily. First consider the forward transfer admittance, y_{21B}. The time-independent component of the current is $-I_p = -\alpha_0 I_E$, where the negative sign arises from the definition of the output-current direction in Fig. 3. The time-dependent component of the current must arise from the charge leaving through the collector, i.e. $Q_s - Q_e$. We can write y_{21B} as

$$y_{21B} = \left.\frac{\partial I_c}{\partial V_e}\right|_{dV_c=0} = \frac{q}{kT}[-\alpha_0 I_E + j\omega(Q_s - Q_e)] \tag{12}$$

The reverse transfer admittance, y_{12B}, is immediately obtained from (12) using (6).

$$y_{12B} = \left.\frac{\partial I_e}{\partial V_c}\right|_{dV_e=0} = \frac{1}{K_c}\frac{q}{kT}[-\alpha_0 I_E + j\omega(Q_s - Q_e)] \tag{13}$$

The remaining output admittance, y_{22B}, employs a current with a time-independent portion $I_p = \alpha_0 I_E$. The time-dependent portion arises from a charge Q_{er}. Q_{er} differs from Q_e only in that the sign of η is opposite, i.e. the direction of the field is reversed when looking into the collector instead of the emitter. Using (8), we may write

$$y_{22B} = \left.\frac{\partial I_c}{\partial V_c}\right|_{dV_e=0} = \frac{1}{K_c}\frac{q}{kT}(\alpha_0 I_E + j\omega Q_{er}) \tag{14}$$

We have derived y_{22B} by inspection, but the validity of using Q_{er} is shown in Appendix III.

The admittances have thus far been written in terms of K_c, Q_s and Q_e in order to allow insight into the origins of the admittances. Q_s is given in (5), an analytical form for Q_e has been found in Appendix II, and K_c is derived in Appendix IV. By using these expressions the admittances can be written in terms of the built-in field. Table 1 lists the admittances in general form separated simply into real and imaginary components. The simplified expressions for zero field (diffusion transistor), and higher aiding fields (drift transistor), are also given in the Table. The results for the diffusion transistor agree with GIACOLETTO[7]. The expressions given in the Table are valid as long as

$$\omega < \frac{1}{\tau_B} = \frac{I_p}{Q_s} = \frac{D}{W^2}\frac{\eta^2}{\eta - 1 + \exp(-\eta)} \tag{15}$$

where τ_B is the transit time of holes across the base region. In today's drift transistors, τ_B can be as small as 0·5 mμ sec. For such a transistor the admittance expressions hold for frequencies less than 320 Mc/s. It is apparent from our model that, when the frequency is comparable to $1/\tau_B$, charging and discharging of the base region is not completely accomplished during a cycle of the applied voltage.

The following Table contains an interesting property of the admittances. Only four of the admittance components depend on the *direction* of the field.

	real part	im. part
y_{11B}	indep.	dep.
y_{12B}	dep.	dep.
y_{21B}	indep.	indep.
y_{22B}	dep.	indep.

An actual transistor has additional components which were not taken into account in the admittances listed in Table 1. Fig. 5 shows the ideal transistor, commonly called "intrinsic", plus the emitter space-charge capacitance C_E, the collector space-charge capacitance C_c and the base resistance

Table 1

	General	Diffusion transistor	Drift transistor $\eta > 3$
y_{11B}	$\frac{qI_E}{kT}\left(1+j\omega\alpha_0\frac{W^2}{D}\frac{1-e^{-\eta}}{\eta}\frac{(\sinh\eta)/\eta-1}{\cosh\eta-1}\right)$	$\frac{qI_E}{kT}\left(1+j\omega\alpha_0\frac{W^2}{3D}\right)$	$\frac{qI_E}{kT}\left(1+j\omega\alpha_0\frac{W^2}{D}\frac{1-2\eta e^{-\eta}}{\eta^2}\right)$
y_{21B}	$\frac{qI_E}{kT}\alpha_0\left[-1+j\omega\frac{W^2}{D}\left(\frac{1}{\eta}\frac{1+e^{-\eta}}{1-e^{-\eta}}-\frac{2}{\eta^2}\right)\right]$	$\frac{qI_E}{kT}\alpha_0\left(-1+j\omega\frac{W^2}{6D}\right)$	$\frac{qI_E}{kT}\alpha_0\left(-1+j\omega\frac{W^2}{D}\frac{\eta-2}{\eta^2}\right)$
y_{12B}	$\alpha_0\frac{I_E}{W}\frac{dW}{dV_c}\frac{\eta}{e^{\eta}-1}\left[-1+j\omega\frac{W^2}{D}\left(\frac{1}{\eta}\frac{1+e^{-\eta}}{1-e^{-\eta}}-\frac{2}{\eta^2}\right)\right]$	$\alpha_0\frac{I_E}{W}\frac{dW}{dV_c}\left(-1+j\omega\frac{W^2}{6D}\right)$	$\alpha_0\frac{I_E}{W}\frac{dW}{dV_c}\left[-\eta e^{-\eta}+j\omega\frac{W^2}{D}e^{-\eta}\left(1-\frac{2}{\eta}\right)\right]$
y_{22B}	$\alpha_0\frac{I_E}{W}\frac{dW}{dV_c}\left(\frac{\eta}{e^{\eta}-1}+j\omega\frac{W^2}{D}\frac{(\sinh\eta)/\eta-1}{\cosh\eta-1}\right)$	$\alpha_0\frac{I_E}{W}\frac{dW}{dV_c}\left(1+j\omega\frac{W^2}{3D}\right)$	$\alpha_0\frac{I_E}{W}\frac{dW}{dV_c}\left(\eta e^{-\eta}+j\omega\frac{W^2}{D}\frac{1-2\eta e^{-\eta}}{\eta}\right)$

r_b. In order to obtain a completely descriptive set of parameters, these additional elements must be combined with the intrinsic admittances already derived. Inclusion of the two capacitances is very simple; $j\omega C_E$ should be added to y_{11B}, and $j\omega C_c$ to y_{22B}. When these two steps are done only the base resistance remains.

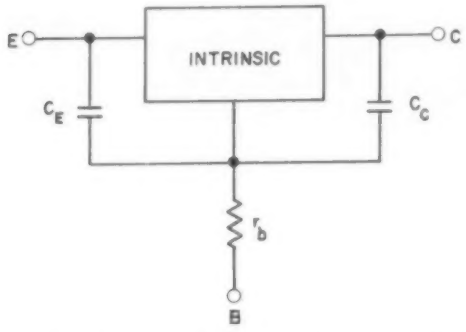


FIG. 5. The base resistance appears in series with the base lead. The junction capacitances shunt the emitter and collector diodes.

The admittance parameters completely define the transistor for small-signal operation. Any other four-pole parameter for any configuration can be obtained from these four admittances. For example, the current gain α is obtained from (11) and (12) for low frequencies as

$$\alpha = -\frac{y_{21B}}{y_{11B}} \simeq \alpha_0\left(1-j\omega\frac{Q_s}{I_p}\right) = \alpha_0(1-j\omega\tau_B) \tag{16}$$

The phase shift of α beginning at low frequencies is due to the stored charge.

Whereas the common base configuration allows us to use a simple method for finding the admittances, transistors are often used in the common-emitter configuration. Transformation of the admittances from common-base to common-emitter is quite simple.

$$\left.\begin{aligned} y_{11E} &= y_{11B}+y_{21B}+y_{12B}+y_{22B} \\ y_{21E} &= -y_{21B}-y_{22B} \\ y_{12E} &= y_{12B}-y_{22B} \\ y_{22E} &= y_{22B} \end{aligned}\right\} \tag{17}$$

Note that y_{12B} and y_{22B} are proportional to dW/dV_c, or to $1/K_c$. Since this multiplier is quite small in practical transistors, these terms can often be neglected when they appear additive to y_{11B} or y_{21B}. This is the case for y_{11E} and y_{21E}. The current gain in common-emitter is

$$\beta = \frac{y_{21E}}{y_{11E}} = \frac{\alpha-y_{22B}/y_{11B}}{1-\alpha+(y_{12B}+y_{22B})/y_{11B}} \tag{18}$$

Apparently, the commonly used $\alpha/(1-\alpha)$ expression for β is not exact. However, the additional terms in (18) are proportional to dW/dV_c, as was just mentioned above. Therefore, in most cases these terms can be neglected. Then, for low frequencies (18) leads to

$$\beta \simeq \beta_0\left(1-j\omega\beta_0\frac{Q_s}{I_p}\right) = \beta_0(1-j\omega\beta_0\tau_B) \tag{19}$$

As can be seen, the phase shift of β beginning at low frequencies is larger than the phase shift of α by the factor β_0. If one combines the junction capacitances with the admittance expressions in Table 1, a fairly accurate frequency dependence is obtained. (The base resistance plays no role in β.)

5. DIFFUSION CAPACITANCES

In general, a diffusion capacitance results from a variation of the stored charge caused by a variation of an applied voltage. It is customary to speak of emitter and collector diffusion capacitances. These capacitances are defined in the common base configuration with the remaining terminal shorted to the base. These definitions happen to coincide with the definitions of y_{11B} and y_{22B}. The emitter diffusion capacitance is therefore

$$\begin{aligned} C_{DE} &= \left.\frac{dQ_e}{dV_e}\right|_{dV_c=0} = \frac{qQ_e}{kT} \\ &= \alpha_0\frac{qI_E}{kT}\frac{W^2}{D}\frac{1-e^{-\eta}}{\eta}\frac{(\sinh\eta)/\eta-1}{\cosh\eta-1} \end{aligned} \tag{20}$$

This capacitance is plotted in normalized form as a

function of η in Fig. 6. We find for a homogeneously doped base region, ($\eta = 0$),

$$C_{DE}(\text{diff.}) = \alpha_0 \frac{qI_E}{kT} \frac{W^2}{3D} \tag{21}$$

If η is large ($\eta \sim 5$), (20) reduces to the approximation given by KRÖMER[8].

$$C_{DE} \to \alpha_0 \frac{qI_E}{kT} \frac{W^2}{D} \frac{1}{\eta^2} \tag{22}$$

This approximation does not cover the whole practical range of drift transistors. The imaginary part of y_{11B}, as given in Table 1 for $\eta > 3$, applies for drift transistors.

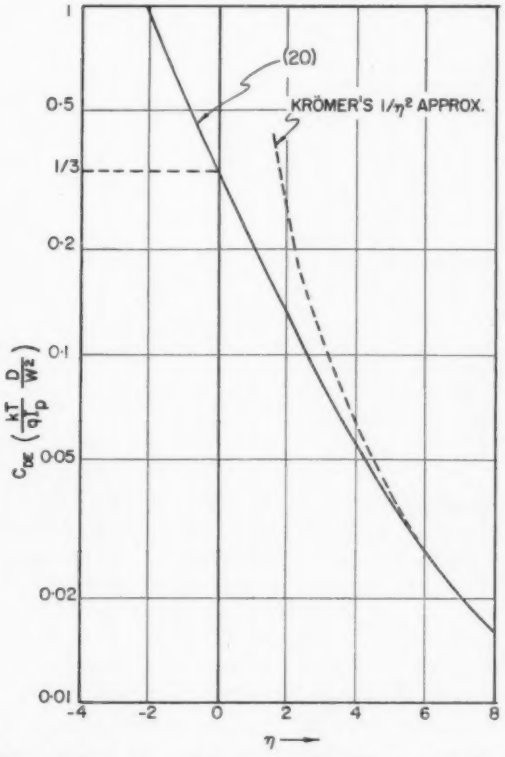


FIG. 6. The emitter diffusion capacitance as a function of the constant built-in field.

The collector diffusion capacitance is defined as the charge moving reversibly through the collector for a variation in the collector voltage.

$$C_{DC} = \left.\frac{dQ_{er}}{dV_c}\right|_{dV_e=0} = \frac{1}{K_c} \frac{qQ_{er}}{kT}$$
$$= \alpha_0 \frac{dW}{dV_C} \frac{I_E W}{D} \frac{(\sinh \eta)/\eta - 1}{\cosh \eta - 1} \tag{23}$$

C_{DC} is also the imaginary part of y_{22B}. It is plotted in a normalized form in Fig. 7. For $\eta = 0$, (23) reduces to

$$C_{DC}(\text{diff.}) = \frac{dW}{dV_c} \alpha_0 \frac{I_E W}{3D} \tag{24}$$

For large values of η, (23) agrees with the approximation given by KRÖMER[8]

$$C_{DC} \to \alpha_0 \frac{dW}{dV_c} \frac{I_E W}{\eta D} \tag{25}$$

This approximation is also plotted in Fig. 7. The collector diffusion capacitance does not depend on the direction of the field. Again, an approximation for a wider range of η is available from Table 1.

Measurement of diffusion capacitance in common base is difficult for several reasons. First of all, the parallel resistance, kT/qI_E, is quite small for moderate currents. Secondly, the base resistance introduces an effective inductance which overcomes the diffusion capacitance. (This effective inductance can be calculated from the admittances.) Also, at high currents other inductive effects[9] appear.

The diffusion capacitance can supply important information about the transistor, and, if it is measured in the common-emitter configuration, much more relevant data are available. Consider that the inductive effect modifies y_{11B} in the following fashion

$$y_{11B} = \frac{q}{kT}(I_E + j\omega Q_e - j\omega Q_i) \tag{26}$$

The inductance introduces some negative imaginary term, $-j\omega Q_i$, which lowers C_{DE}. This means less charge leaves through the emitter and more through the collector during discharge. Therefore, the same term must be added to y_{21B}.

$$y_{21B} = \frac{q}{kT}(-\alpha_0 I_E + j\omega[Q_s - Q_e] + j\omega Q_i) \tag{27}$$

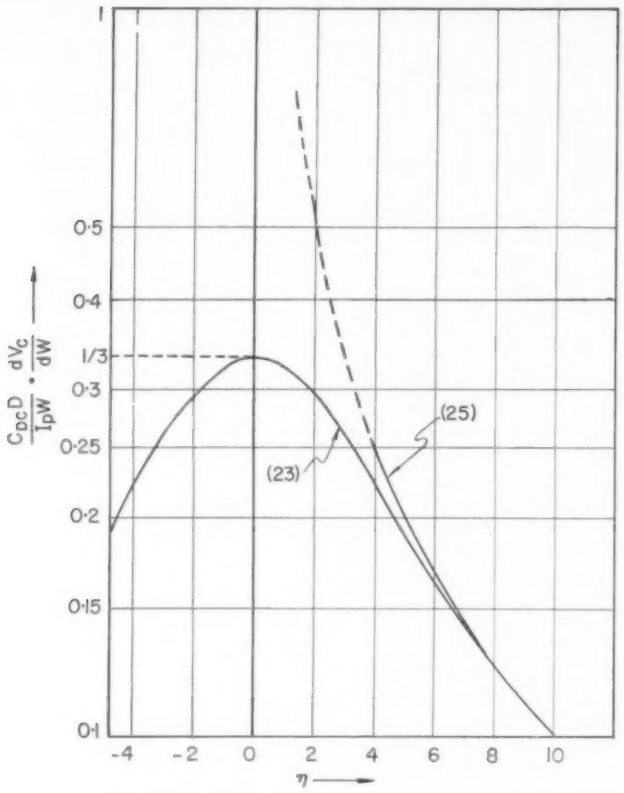


FIG. 7. The collector diffusion capacitance as a function of the constant built-in field. Note that the collector diffusion capacitance is independent of the direction of the built-in field.

Inductive effects do not appear in y_{11E} in the common-emitter configuration. Ignoring the base resistance for the moment, and including the junction capacitances, from (17)

$$\begin{aligned} y_{11E} &= y_{11B}+j\omega C_E+y_{21B}+y_{12B}+y_{22B}+j\omega C_c \\ &\simeq y_{11B}+y_{21B}+j\omega(C_E+C_c) \\ &= \frac{q}{kT}\left(\frac{I_E}{1+\beta_0}+j\omega Q_s\right)+j\omega(C_E+C_c) \end{aligned} \tag{28}$$

Another diffusion capacitance C_D is defined from y_{11E}, which is the total charge divided by the thermal voltage. This holds for any transistor. It is convenient to use y_{11E} for measurement since inductive effects are not present and the parallel resistance is $(1+\beta_0)$ times larger than in the common-base configuration. One can see that (28) presents a means of obtaining τ_B, an important circuit-independent parameter of the transistor, (15). A simple measuring circuit is shown in Fig. 8, and a measurement of C_D made on a drift transistor is shown in Fig. 9. The measured capacitance is $C_m = C_D+C_E+C_c$.

Measurement of C_D allows monitoring the stored charge, and therefore τ_B. For high current levels, the high built-in field in a drift transistor will be "washed out" by the high minority-carrier density. Then one would expect that Q_s would increase faster than linearly with the current[10] and that C_D would also. However, in all practical cases the measured capacitance will

tend to saturate in value owing to the base resistance. The equivalent circuit for the input admittance is shown in Fig. 10. The proportionality factor between the measured and actual capacitances is unity for low currents, because $\beta_0 kT/qI_E$ is large. For higher currents the factor is lowered, especially because β also deteriorates. Measurement of the high-current behavior is difficult since this factor must be taken into account.

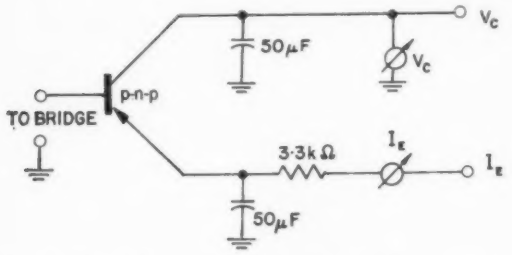


FIG. 8. Circuit for measurement of C_m in the common-emitter configuration. The bridge provides a bias path for the base.

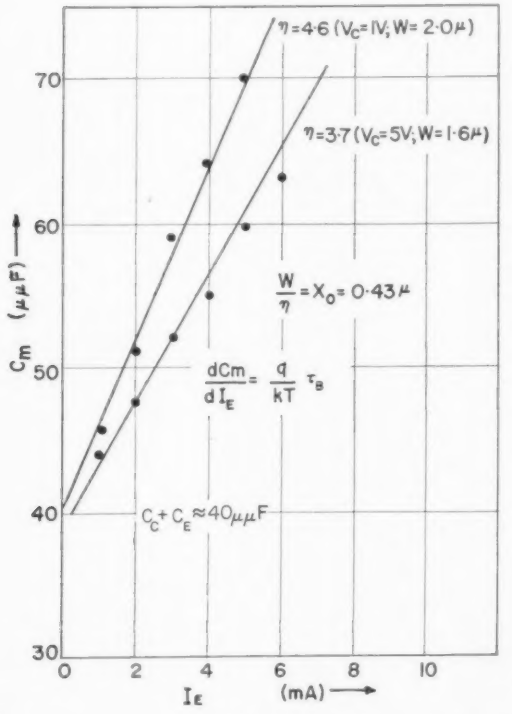
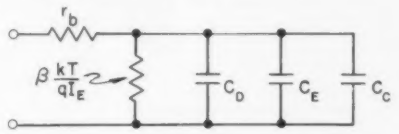


FIG. 9. A measurement of C_m for a drift transistor at two collector voltages.



$$C_m = \left(\frac{\beta\frac{kT}{qI_E}}{\beta\frac{kT}{qI_E} + r_b}\right)^2 (C_D + C_E + C_C)$$

FIG. 10. The equivalent circuit for the common-emitter input admittance. The resistive divider will lower the value of C_m for high currents.

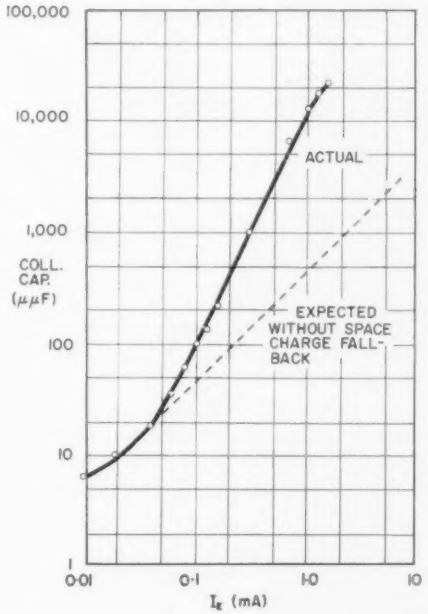


FIG. 11. A measurement of C_{DC} for an abrupt-junction drift transistor. The extreme increase with current is due to "compression" of the collector space charge at low voltages.

The collector capacitance measured in either configuration is $C_{DC}+C_c$. One would expect that the measured collector capacitance would increase linearly with the current. However, as the current is increased, the collector space-charge width does not remain constant.[11] For abrupt collector junction, the space charge narrows, giving rise to a large increase in C_c, as Fig. 11 shows.

T

Apparently the variation of space-charge thickness with current can be monitored well.

Although we have emphasized the diffusion capacitances because of their practical importance, many other properties of a transistor can be described by means of the simple admittance parameters determined here.

Acknowledgements—The authors wish to thank A. SLOBODSKOY of this laboratory for his assistance, and K. ARBENZ for the computation.

REFERENCES

1. W. SHOCKLEY, *Bell Syst. Tech. J.* **28,** 435 (1949).
2. H. KRÖMER, *Arch. Elektr. Übertr.* **8,** 226 (1954).
3. R. BEAUFOY and J. J. SPARKES, *A.T.E.J.* **13,** 320 (1957) and *Proc. I.R.E.* **45,** 1740 (1957).
4. J. LINDMAYER and C. WRIGLEY, *Solid-State Electron.* **2,** 247 (1961).
5. J. ZAWELS, *RCA Rev.* **16,** 366 (1955).
6. J. LINDMAYER and C. WRIGLEY, *Semicond. Prod.* **4,** 41 (1961).
7. L. J. GIACOLETTO, *RCA Rev.* **15,** 529 (1954).
8. H. KRÖMER, The drift transistor, *Transistors I* 209, 211, RCA Labs., Princeton, N. J. (1956).
9. T. E. FIRLE and O. E. HAYES, *I.R.E. Trans.* **ED6,** 330 (1959).
10. J. LINDMAYER and C. WRIGLEY, *Solid-State Electron.* **2,** 79 (1961).
11. C. T. KIRK, *M. I. T. Lincoln Lab. Rep.* 53G–0056 (1961).

APPENDIX I

We can obtain an expression for the built-in electric field from the electron-current expression. Neglecting the electron current, we can write

$$I_n = 0 = qAD_n n' + qA\mu_n E_b n \tag{I.1}$$

Utilizing the Einstein relation ($D = \mu kT/q$) we can solve (I.1) for the field and write

$$E_b = -\frac{kT}{q}\frac{n'}{n} \tag{I.2}$$

Space-charge neutrality in the base region requires that

$$n - N - p = 0 \tag{I.3}$$

Setting (I.3) into (I.2),

$$E_b = -\frac{kT}{q}\frac{(N+p)'}{N+p} \tag{I.4}$$

For not-too-high currents, $p \ll N$, and, therefore, making use of (1) in the text for exponential N,

$$E_b = -\frac{kT}{q}\frac{N'}{N} \to \frac{kT}{qW}\eta \tag{I.5}$$

APPENDIX II

The charges giving rise to a capacitive current by being reclaimable through the emitter can be calculated from the continuity equation

$$\frac{\partial p}{\partial t} = D\frac{\partial^2 p}{\partial x^2} - \frac{\eta}{W}D\frac{\partial p}{\partial x} \tag{II.1}$$

because $q_A dp/dt = -\text{div. } I_p$. I_p is given in (3). The steady-state distribution of holes is given in (4). At $t = 0$ the concentration of holes at the emitter is instantaneously dropped to zero by dropping the emitter voltage. The solution is

$$p(x,t) = \frac{IpW}{qAD}\frac{1-e^{-\eta}}{\eta}2\pi \exp(\eta x/2W)\sum_{m=1}^{\infty} \exp\{-D/W^2[m^2\pi^2+(\eta/2)^2]t\}\frac{m \sin m\pi x/W}{m^2\pi^2+(\eta/2)^2} \tag{II.2}$$

The discharge current crossing the emitter can be calculated from the slope of the hole distribution at $x = 0$. Since the concentration is zero at $x = 0$, only diffusion current should be considered.

$$I_R = -qADp'(x=0) \tag{II.3}$$

where the R subscript refers to the discharge current coming out of the emitter. And, finally, the charge reclaimable through the emitter is

$$Q_e = \int_0^\infty I_R dt = I_p\frac{W^2}{D}\frac{1-e^{-\eta}}{\eta} \times 2\sum_{m=1}^{\infty}\left[\frac{1}{m\pi + \dfrac{(\eta/2)^2}{m\pi}}\right]^2 \tag{II.4}$$

It is not convenient to have a summation in the final form. With some mathematical and physical manipulations, it is possible to find an analytical form. However, the exact equality could not be proven directly. It was

found that

$$\sum_{m=1}^{\infty}\left[\frac{1}{m\pi+\dfrac{(\eta/2)^2}{m\pi}}\right]^2=\tfrac{1}{2}\frac{(\sinh\eta)/\eta-1}{\cosh\eta-1} \tag{II.5}$$

The summation was computed and the results are compared with the analytical form in Table 2. Apparently the summation approaches the analytical form proposed, and the equality in (II.5) may be assumed to be correct. Rewriting the reclaimable charge in (II.4) we obtain

$$Q_e=\frac{I_pW^2}{D}\frac{1-e^{-\eta}}{\eta}\frac{(\sinh\eta)/\eta-1}{\cosh\eta-1} \tag{II.6}$$

Since we are dealing with a linear system, any arbitrarily smaller step in the emitter voltage will simply result in a proportionately smaller Q_e; i.e. dQ_e is proportional to dI_p, which represents the change in the steady-state hole current.

Table 2†

η	Σ	Number of terms computed (m)	Analytical form
0	0·16626	247	0·16666
2	0·14685	263	0·14724
4	0·11032	304	0·11066
6	0·08097	354	0·08125
8	0·06195	405	0·06221
10	0·04973	452	0·04996

† Comparison of computed summation and proposed analytical form. The computation was stopped when the m term was 0·1 per cent or less of the terms already collected.

APPENDIX III

When the collector voltage is changed, a change in the collector space-charge width occurs. When this dW change is introduced at the collector, the charge change can be obtained by rewriting (4) as

$$p=p_e\frac{1-\exp[-\eta(1-x/W)]}{1-e^{-\eta}} \tag{III.1}$$

where p_e is the concentration at $x=0$ and is constant, since the emitter voltage is unchanged due to the definition of y_{22B}. Therefore, the charge change is as shown qualitatively in Fig. 12.

The change in the hole concentration for a dW change in basewidth is obtained by differentiating (III.1) with respect to W.

$$dp=\frac{dW}{W}p_e\eta\frac{\exp[-\eta(1-x/W)]-e^{-\eta}}{(1-e^{-\eta})^2}$$

$$=-\frac{I_p}{qAD}dW\frac{\eta}{e^{\eta}-1}\frac{1-\exp(\eta x/W)}{\eta} \tag{III.2}$$

Now, if we invert the x-axis (defining $X=W-x$) we move the boundary, where we will observe the discharge current, to $X=0$.

$$dp=\left(\frac{I_p\,dW}{qAD}\frac{\eta}{e^{\eta}-1}\right)\frac{1-\exp[\eta(1-X/W)]}{-\eta} \tag{III.3}$$

The multiplier in (III.3) is constant with respect to X. We can now see that (III.3) is similar to (4); the only difference is the opposite sign on η.

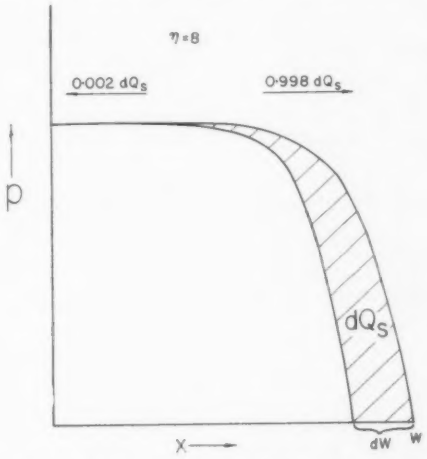


FIG. 12. For a change in the collector voltage, a dW change in the basewidth occurs. The excess charge leaves almost entirely through the collector for high built-in fields.

The charge passing through the collector for a dW change can be obtained without solving the continuity equation again. The only change is in the initial distribution of holes, which can be taken into account by using the new constant and the opposite sign for η. The charge reclaimable for a dW decrease is

$$dQ_c=\frac{I_pW}{D}dW\frac{(\sinh\eta)/\eta-1}{\cosh\eta-1} \tag{III.4}$$

According to the formulation of y_{22B}, the imaginary part must be $j\omega dQ_c/dV_c$. In (14) we found by inspection that the imaginary part is $j\omega qQ_{er}/K_c kT$. (III.4) can be written in terms of Q_{er}, [invert the sign of η in (II.6)], and K_c (as derived in Appendix IV). The result of rewriting is

$$\frac{dQ_c}{dV_c} = \frac{1}{K_c}\frac{qQ_{er}}{kT} \tag{III.5}$$

which proves the validity of writing (14) by inspection.

APPENDIX IV

Since K_c is a real number for all transistors at all frequencies, it is worthwhile to calculate it in general. Inserting the hole current into (8) we can write

$$K_c = \frac{dI_p}{dV_e}\bigg|_{dV_e=0}\frac{dV_c}{dW}\frac{dW}{dI_p}\bigg|_{dV_e=0} \tag{IV.1}$$

While dI_p/dV_e is readily available from (10) and dV_c/dW is a constant of the transistor, dI_p/dW must be calculated.

$$K_c = \frac{qI_p}{kT}\frac{dV_c}{dW}\frac{dW}{dI_p}\bigg|_{dV_e=0} \tag{IV.2}$$

Let us write the hole current in general, making use of the Einstein relation and the field expression in (I.5).

$$I_p = -qADp' + qA\mu E_b p = -qAD\left(p' + \frac{N'}{N}p\right) \tag{IV.3}$$

This is a linear differential equation for p. The general solution, with the boundary condition $p = 0$ at $x = W$, is

$$p = \frac{I_p}{qAD}\frac{\int_x^W N\,dx}{N} \tag{IV.4}$$

The concentration of holes at the emitter boundary is

$$p_e = \frac{I_p}{qAD}\frac{\int_0^W N\,dx}{N_E} \tag{IV.5}$$

We are looking for dI_p/dW with the condition that $dV_e = 0$. Then, $dp_e/dW = 0$ must also apply. Differentiating (IV.5) with respect to W, we find

$$\frac{dp_e}{dW} = 0 = \frac{I_p}{qADN_E}\frac{d}{dW}\int_0^W N\,dx + \frac{dI_p}{dW}\frac{1}{qADN_E}\int_0^W N\,dx \tag{IV.6}$$

or

$$\frac{dI_p}{dW}\bigg|_{dV_e=0} = -I_p\frac{\frac{d}{dW}\int_0^W N\,dx}{\int_0^W N\,dx} \tag{IV.6}$$

The numerator can be simplified. The derivative of the lower limit is obviously zero. The differential of the upper limit is just $N(W)$, i.e. the x-variable in the original N-function is replaced by W. Setting this result into (IV.2) we have

$$K_c = \left|\frac{dV_c}{dW}\right|\frac{q}{kT}\frac{\int_0^W N\,dx}{N(W)} \tag{IV.7}$$

This is the general expression for K_c for any transistor. Specifically for the exponential impurity distribution, (1), we have

$$K_c = \frac{dV_c}{dW}\frac{qW}{kT}\frac{e^{\eta}-1}{\eta} \tag{IV.8}$$

K_c is always a positive large number. dV_c/dW should be regarded as a positive constant in all the above expressions.

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 291–303. Printed in Great Britain

CHARACTERISTICS OF THE SPACE-CHARGE-LIMITED DIELECTRIC DIODE AT VERY HIGH FREQUENCIES*

J. SHAO† and G. T. WRIGHT

Electrical Engineering Department, University of Birmingham, England

(*Received* 10 *June* 1961; *in revised form* 14 *July* 1961)

Abstract—The incremental admittance of the space-charge-limited dielectric diode is calculated; it is shown that for all frequencies this can be represented as a conductance and susceptance in parallel. At very high frequencies, transit time effects reduce the forward conductance to 0·667 of the low frequency value and increase the input capacitance to 1·33 of the low frequency value. At low frequencies such that transit time effects can be neglected the reverse biased capacitance is constant and equal to the geometrical electrode capacitance; the forward biased capacitance is constant and equal to 0·75 of the reverse biased capacitance. There are no negative resistance effects at any frequency.

Experimental measurements have been made on space-charge-limited cadmium sulphide diodes over the frequency range 50–1000 Mc/s. These confirm the theoretical analysis and support the physical model used for description of space-charge-limited current in solids.

Résumé—L'accroissement d'admittance d'une diode diélectrique à charge d'espace limitée est calculé. On démontre qu'on peut représenter cet accroissement par un circuit comprenant une conductance et une susceptance en parallèle. Aux très hautes fréquences, les effets du parcours électronique réduisent la conductance avant à 0,667 de sa valeur à basse fréquence et augmentent la capacité d'entrée à 1,33 de sa valeur à basse fréquence. Aux basses fréquences, les effets du parcours électronique peuvent être ignorés, la capacité polarisée à sens inverse est constante et égale à la capacité géométrique de l'électrode; la capacité polarisée à sens avant est aussi constante et égale à 0,75 de la valeur à sens inverse. A toutes les fréquences, il n'existe pas d'effets de résistance négative.

Des mesures expérimentales ont été prises sur des diodes de sulfure de cadmium à charge d'espace limitée pour une gamme de fréquences allant de 50 à 1000 M/sc. Celles-ci confirment l'analyse théorique et prouve le modèle physique employé pour décrire le courant à charge d'espace limitée dans les solides.

Zusammenfassung—Der zusätzliche Scheinwiderstand der dielektrischen Diode mit begrenzter Raumladung wird berechnet. Er lässt sich bei allen Frequenzen als Parallelanordnung einer Konduktanz und einer Suszeptanz darstellen. Bei sehr hohen Frequenzen reduzieren Durchgangszeit-Effekte die Konduktanz in Vorwärtsrichtung auf 0,667 des Wertes bei niedriger Frequenz, und sie erhöhen die Eingangskapazität auf 1,33 ihres Wertes bei niedriger Frequenz. Bei Niederfrequenzen, die eine Vernachlässigung der Durchgangszeit-Effekte zulassen, ist die in Rückwärtsrichtung vorgespannte Kapazität konstant und gleich der geometrischen Elektrodenkapazität. Die in Vorwärtsrichtung vorgespannte Kapazität ist konstant und gleich 0,75 der in Rückwärtsrichtung vorgespannten Kapazität. Bei keiner Frequenz ergeben sich negative Widerstände.

Messungen an derartigen Dioden aus Cadmiumsulfid wurden im Frequenzbereich 50–1000 MHz vorgenommen. Sie bestätigen die theoretische Ableitung und unterstützen das zur Beschriebung des in Festkörpern auftretenden Stromes mit begrenzter Raumladung benutzte physikalische Modell.

1. INTRODUCTION

A COMPARISON of the diffusion-limited semiconductor junction diode and the space-charge-limited dielectric diode suggests that the former is basically suited to power applications whereas the latter is basically suited to electronics applications.[1] This follows because the semiconductor diode is capable of operation at very high current densities but is noisy and temperature sensitive,

*This paper was submitted by J. SHAO in partial fulfillment of the requirements for the M.Sc. degree of Birmingham University.

†Present address: School of Applied Physics, Harvard University, Cambridge, Mass.

whereas the dielectric diode is capable of operating up to very high frequencies and should possess a low noise figure and be relatively temperature insensitive.

This paper focuses interest on the very high-frequency characteristics of the space-charge-limited (SCL) dielectric diode. The small-signal theory of the diode is worked out for frequencies at which electron transit times are significant and experimental data are presented to confirm the essential validity of the physical model used to describe the mechanisms of SCL current in solids.

NOTATION

B_1 first-order susceptance
C geometric capacitance of diode per unit area
C_0 geometric capacitance of non-injecting areas of cathode
C_s geometric capacitance of injecting areas of cathode
E electric field intensity
G_1 first-order conductance
I total current
I_0 steady-current component
I_1 amplitude of first-order variation in current
R diode forward resistance
T transit time for the distance x
T_0 steady transit time component
T_a transit time from virtual cathode to anode
T_1 amplitude of first-order variation in transit time
V potential difference between virtual cathode and anode
V_0 steady applied voltage
V_1 amplitude of first-order variation in applied voltage
V_a applied voltage
V_t threshold voltage
Y_1 first-order incremental admittance
d diode interelectrode spacing
g low-frequency incremental conductance
p $j\omega$
r low-frequency incremental resistance
t_c time at which electrons leave virtual cathode
v velocity of electrons
v_0 steady-velocity component
v_1 amplitude of first-order variation in electron velocity
x distance from cathode
ρ charge density
μ electron mobility
ϵ permittivity
θ transit angle $= \omega T_a$

2. THEORY

The small-signal theory of transit-time effects for electrons in vacuum tubes has been worked out by BENHAM[2] and LLEWELLYN[3]. The method of analysis used here is based on the approach of these authors but is modified where necessary, as we are here discussing electron motion in solids and not in vacuum. The one-dimensional, one-carrier (electron) case is treated and equations are derived describing the SCL current in the dielectric diode when time-varying voltages are applied. The case of a small sinusoidal voltage superimposed on a large steady voltage is examined in detail and a solution is obtained which expresses the incremental admittance of the diode, a readily measurable quantity, in terms of the transit time of the electrons.

A. *Basic equations*

The equations which describe current in the dielectric diode when time-varying voltages are applied may be written down readily.

The basic expression which describes the total current density as the sum of conduction and displacement currents is Maxwell's equation:

$$I = \rho v + \epsilon \frac{\partial E}{\partial t} \tag{1}$$

This basic equation is of quite general validity. However, when current occurs under SCL conditions, the electric-field intensity and the electron space-charge density are dependent on each other. For SCL current in vacuum or for SCL current in a trap-compensated solid this is expressed by Poisson's equation in the form

$$\epsilon \frac{\partial E}{\partial x} = \rho \tag{2}$$

The last of the basic equations may be termed the equation of motion of the electron space-charge. This is quite different for the two cases of current in vacuum and in a solid. In vacuum, electrons move from cathode to anode with free acceleration and their velocity v after travelling a distance x from the cathode and falling through a potential difference V is $v = (2\ \mathrm{eV}/m)^{1/2}$. In this expression the small but finite thermal velocities of emission from the cathode are neglected. This is justified because the energy gained by the space charge from the electric field is much greater than its thermal energy over the greater part of the space from cathode to anode. In the solid, however, the situation is far more complex. The mean thermal energy of the space charge is greater than the energy gained from the electric field, for the latter is continuously being dissipated by collisions

with the crystal lattice. A consequence of this is that the partial pressure of the mobile electrons exerts a force on the space charge which is comparable with that exerted by the applied electric field over an appreciable part of the cathode-to-anode space. This has been analysed in detail[4] and the essential result is that a "virtual cathode" is formed, as in the case of the vacuum diode, but is now far more pronounced and is formed at a much greater relative distance from the cathode itself. By considering electron emission to take place from the virtual cathode, the influence of the diffusion-current tendency is taken into account and one may then consider that, from the virtual cathode to the anode, current occurs by electron drift. Thus for the case of SCL current in solids the equation of motion may be written

$$v = \mu E \tag{3}$$

This equation implies that the mean time between collisions of the space-charge electrons with the crystal lattice is much shorter than the period of the time-varying applied voltage. This will always be the case in a device of practical dimensions, for each space-charge electron will make many collisions with the crystal lattice in moving from cathode to anode. This enables us to neglect the inertia effects of the electrons in transit and to regard their response to the applied electric field to be instantaneous.

The very high-frequency characteristics of the dielectric diode are described by equations (1–3). These may be expressed in a much simpler and more compact form.

Consider the movement of an electron between two planes spaced distance dx apart at a distance x from the virtual cathode; the electron will require a time dt to cross this space. The time rate of change of electric field experienced by this electron is

$$\frac{dE}{dt} = \frac{\partial E}{\partial x}\frac{dx}{dt} + \frac{\partial E}{\partial t} \tag{4}$$

Using equation (2) we may write equation (4) as

$$\frac{dE}{dt} = \frac{\rho v}{\epsilon} + \frac{\partial E}{\partial t} \tag{5}$$

Substituting this into equation (1) we have

$$I = \epsilon\frac{dE}{dt} \tag{6}$$

With the aid of equation (3) this becomes

$$I = \frac{\epsilon}{\mu}\frac{dv}{dt} \tag{7}$$

This result shows that the total current I is a function of time only; at a given instant the current density is the same through any plane between virtual cathode and anode.

B. *Small-signal theory of the dielectric diode*

Our concern is to describe the response of the SCL dielectric diode to an applied time-varying voltage; in particular we are interested in the situation when a small-amplitude sinusoidal voltage is superimposed on to a large and steady voltage.

Since current is a function of time only and since we are considering only small-signal conditions, we may write

$$I = I_0 + I_1 e^{pt} \tag{8}$$

where I_0 is the steady-current component and I_1 is the amplitude of the fundamental frequency; as usual $p = j\omega$. Using equation (7) we have

$$\frac{dv}{dt} = \frac{\mu}{\epsilon}(I_0 + I_1 e^{pt}) \tag{9}$$

which may be integrated to give

$$v = \frac{\mu}{\epsilon}\left(I_0 t + \frac{I_1 e^{pt}}{p}\right) + \text{const.} \tag{10}$$

Under SCL conditions the electric field is zero at the virtual cathode and we may consider the velocity of the electrons to be zero at this point. Thus if t_c is the time at which the electrons left the virtual cathode we have $v = 0$ when $t = t_c$. Using this initial condition equation (10) becomes

$$v = \frac{\mu}{\epsilon}\left[I_0(t - t_c) + \frac{I_1 e^{pt}}{p} - \frac{I_1 e^{pt_c}}{p}\right] \tag{11}$$

Integrating again and using the initial condition that $x = 0$ when $t = t_c$ we have

$$x = \frac{\mu}{\epsilon}\left[\frac{I_0}{2}(t - t_c)^2 + \frac{I_1 e^{pt}}{p^2} - \frac{I_1 e^{pt_c}}{p^2} - (t - t_c)\frac{I_1 e^{pt_c}}{p}\right] \tag{12}$$

The time interval $(t-t_c)$ is the time taken by the electrons to travel the distance x. In the steady-current case when the applied voltage is constant, I_1 is zero and we may write

$$x = \frac{\mu}{\epsilon}\left(\frac{I_0 T_0^2}{2}\right) \tag{13}$$

where $T_0 = (t-t_c)$ and is the transit time for the distance x.

When the time-varying voltage is present, the transit time will be time varying and we may write

$$t-t_c = T_0 + T_1 e^{pt} \tag{14}$$

where T_1 is the amplitude of the first-order time variation. This expression may be used to eliminate t_c in equations (11) and (12). Using Taylor's expansion for $\exp[p(t-T_0-T_1e^{pt})]$ and retaining only first-order terms we obtain

$$v = \frac{\mu}{\epsilon}\left[I_0(T_0 + TIe^{pt}) + \frac{I_1 e^{pt}}{p} - \frac{I_1 e^{p(t-T_0)}}{p}\right] \tag{15}$$

$$x = \frac{\mu}{\epsilon}\left[\frac{I_0 T_0^2}{2} + I_0 T_0 T_1 e^{pt} + \frac{I_1 e^{pt}}{p^2} - \frac{I_1 \exp[p(t-T_0)]}{p^2} - \frac{I_1 T_0 e^{p(t-T_0)}}{p}\right] \tag{16}$$

Equations (13) and (16) may now be used to obtain an expression for T_1, the first-order time variation in transit time. Substituting this into equation (15) we obtain

$$v = \frac{\mu}{\epsilon}(I_0 T_0) + \frac{\mu}{\epsilon}\left(\frac{1}{T_0}\right)\left[\frac{I_1 T_0}{p} - \frac{I_1}{p^2} + \frac{I_1 e^{-pT_0}}{p^2}\right] e^{pt} \tag{17}$$

This result gives the velocity of the electron space charge at any time t and at any distance x (although x is represented here by T_0).

To separate the velocity into its steady and time-varying components we write

$$v = v_0 + v_1 e^{pt} \tag{18}$$

and equate this to equation (17) to obtain

$$v_0 = \frac{\mu I_0 T_0}{\epsilon} \tag{19}$$

$$v_1 = \frac{\mu}{\epsilon}\left(\frac{1}{T_0}\right)\left[\frac{I_1 T_0}{p} - \frac{I_1}{p^2} + \frac{I_1 e^{-pT_0}}{p^2}\right] \tag{20}$$

Now the electric field E is everywhere proportional to v as expressed by equation (3); consequently we have in effect found the electric field E in terms of the current I. In order to obtain the final solution describing current in terms of applied voltage, we may integrate the electric-field intensity E over the electrode spacing of the diode to obtain the applied voltage V as

$$V = -\int_0^d E\partial x = -\frac{1}{\mu}\int_0^d v\partial x \tag{21}$$

In order to evaluate this integral we make use of the fact that velocity v has been found in terms of transit time T_0 and that a relation between distance x and transit time T_0 has been found which is independent of time. Thus a change of variable from x to T_0 will enable the integration to be carried out.

Using equation (13) we have

$$\partial x = \frac{\mu}{\epsilon} I_0 T_0 \partial T_0 \tag{22}$$

If we now denote the transit time from virtual cathode to anode by T_a, we have from equations (21) and (22) that

$$V = \frac{1}{\epsilon}\int_0^{T_a} [v_0 + v_1 e^{pt}] I_0 T_0 \partial T_0 \tag{23}$$

In order to separate the applied voltage into its steady and time-varying components we write

$$V = V_0 + V_1 e^{pt} \tag{24}$$

Equating this with equation (23) we have

$$V_0 = \frac{1}{\epsilon}\int_0^{T_a} v_0 I_0 T_0 \partial T_0 \tag{25}$$

$$V_1 = \frac{1}{\epsilon}\int_0^{T_a} v_1 I_0 T_0 \partial T_0 \tag{26}$$

(i) *Steady-current relations.* From equation (25) we may obtain the steady-current characteristics of the SCL dielectric diode. Substituting for v_0 from equation (19) we have

$$V_0 = \frac{\mu}{\epsilon^2} I_0^2 \int_0^{T_a} T_0^2 \partial T_0 \tag{27}$$

Integrating this we obtain the applied steady voltage in terms of the electron transit time

$$V_0 = \tfrac{1}{3}\frac{\mu}{\epsilon^2} I_0^2 T_a^3 \tag{28}$$

Using now equation (13) to eliminate the transit time we obtain on re-arranging

$$I_0 = \tfrac{9}{8}(\epsilon\mu)\frac{V_0^2}{d^3} \tag{29}$$

which is the fundamental and well-known Child's law for SCL current in solids.

It is useful to express the ratio of steady current to steady voltage as a "resistance"

$$R = \frac{V_0}{I_0} = \frac{2T_a d}{3\epsilon} = \frac{2T_a}{3C} \tag{30}$$

where $C = \epsilon/d$ and is the geometric capacitance of the diode per unit area.

The incremental resistance r may be obtained from equation (29) as

$$r = \frac{\partial V_0}{\partial I_0} = \frac{V_0}{2I_0} = \frac{R}{2} = \frac{T_a}{3C} \tag{31}$$

(ii) *Incremental admittance.* The incremental admittance may be obtained by using equation (26) to evaluate the first-order time-varying current in terms of the time-varying applied voltage.

Using equation (20) to eliminate v_1 and integrating, we obtain

$$V_1 = \frac{\mu}{\epsilon^2}\frac{I_0 I_1}{p^3}\left[\frac{p^2 T_a^3}{2} - pT_a - e^{-pT_0} + 1\right] \tag{32}$$

Using now the concept of transit angle defined by

$$\theta = \omega T_a \tag{33}$$

we may evaluate the incremental admittance $Y_1 = I_1/V_1 = G_1 + jB_1$. We obtain

$$G_1 = \frac{g\theta^3}{6}\,\frac{\theta - \sin\theta}{(\theta - \sin\theta)^2 + (\theta^2/2 + \cos\theta - 1)^2} \tag{34}$$

$$B_1 = \frac{g\theta^3}{6}\,\frac{\theta^2/2 + \cos\theta - 1}{(\theta - \sin\theta)^2 + (\theta^2/2 + \cos\theta - 1)^2} \tag{35}$$

In these expressions g is the steady-current incremental conductance defined by $g = 1/r$. The frequency dependence of these parameters is illustrated in Fig. 1.

At low and moderate frequencies the period of the applied voltage will be much longer than the electron transit time and we shall have $\theta \ll 1$. Under these conditions the expressions for the incremental conductance and susceptance may be simplified and give

$$G_1 = g; \quad B_1 = g\theta/4 \tag{36}$$

so that the incremental admittance may be written

$$Y_1 = g + j\frac{3\omega C}{4} \tag{37}$$

The significance of this result is to show that at low and moderate frequencies, for which transit time effects may be neglected, the small-signal equivalent circuit for the diode may be regarded as a conductance g in parallel with a capacitance $3C/4$.

At very high frequencies the period of the applied voltage will be comparable with or shorter than the electron transit time, and the full expressions for G_1 and B_1 must be used to evaluate the diode admittance.

3. EXPERIMENTAL

The theoretical analysis which has been worked out in the preceding Section is based on a simplified physical model for the mechanisms of SCL current in solids and applies strictly to the idealized case of vanishingly small signal conditions. The calculations of incremental admittance and transit-time effects therefore refer to a somewhat idealized situation and it is desirable to confirm experimentally that the analysis is indeed applicable to the real case.

This Section describes the fabrication of experimental diodes and the results of measurements of very high-frequency characteristics.

A. *Fabrication of SCL cadmium sulphide diodes*

Electron-transit times in the SCL dielectric diode are very short and the device can therefore operate up to very high frequencies. This means that measurements must be made at high frequencies in order to observe and separate the conductance and susceptance components of the admittance and at still higher frequencies in order to observe transit-time effects. Even using cadmium sulphide crystals, in which the electron mobility is relatively low, transit-time effects are not be be expected until the kilomegacycle frequency range is reached. However, by using coaxial line measurement techniques and by working with a small applied steady voltage so that the transit time should not be too short, it has been found possible to make satisfactory measurements to confirm the essential validity of the theoretical analysis and the physical model on which it is based.

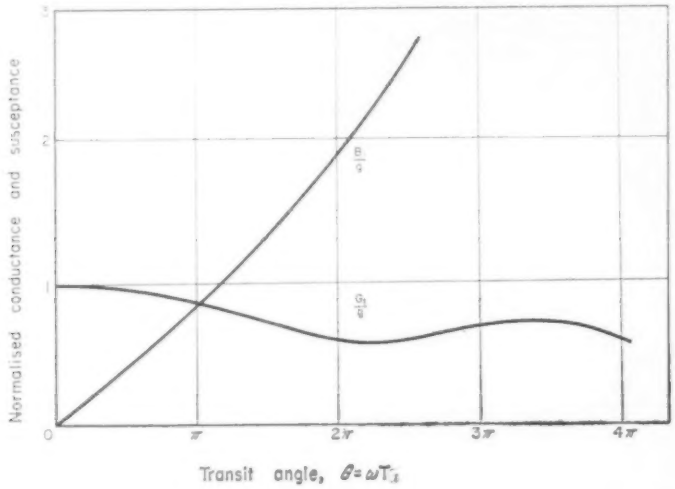


FIG. 1. Dependence of incremental conductance and susceptance upon transit angle calculated from small-signal theory.

For these very high-frequency measurements diodes were made in coaxial form using commercial microwave diode cartridges. The method of construction is illustrated in Fig. 2. The cadmium sulphide crystal was first bonded with indium or tin to the end of a nickel rod; this formed the low-resistance, electron-injecting contact to the crystal.[5] The high-resistance contact was then formed on the other side of the crystal by vacuum evaporation of gold. The nickel rod carrying the crystal was then inserted into the top brass stud, and the top and bottom studs were then pushed into the ceramic body of the cartridge and bonded in place with Araldite. Contact between the evaporated gold anode on the crystal and the bottom brass stud was then made by carefully pressing the crystal against a soft indium dot previously soldered on to the bottom stud.

A large number of diodes have been made in this way and in most cases a square-law dependence of steady current upon voltage was observed. This

is in accordance with the fundamental Child's law derived in the preceding Section.

A typical plot of the square root of steady forward current against applied voltage is shown in Fig. 3. The characteristic shows the usual threshold voltage V_t originating basically in the difference between the electron affinity of the crystal and the work function of the anode metal, which must be equalled by the applied voltage before SCL current begins to occur. It is relevant to note, in this connexion, that the theoretical analysis of the preceding Section is made in terms of the inner potential difference between the virtual cathode and the anode surface of the crystal. This is equal to the applied voltage V_a less the threshold voltage V_t as given by

$$V = V_a - V_t$$

It is apparent from these steady-current observations that trap-compensation in this crystal was not exact; this is indicated by the deviation of the characteristic from the fundamental square-law behaviour for small currents. It is also apparent, however, that, for an applied inner potential greater than about 0·2 V, these residual uncompensated trapping levels are saturated, and for applied voltages greater than this the crystal may be regarded as essentially trap-free. A number of diodes of this type were selected for the high-frequency measurements.

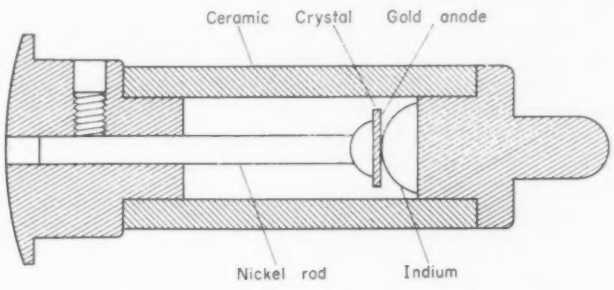


FIG. 2. Illustrating the coaxial construction used for very high-frequency diodes.

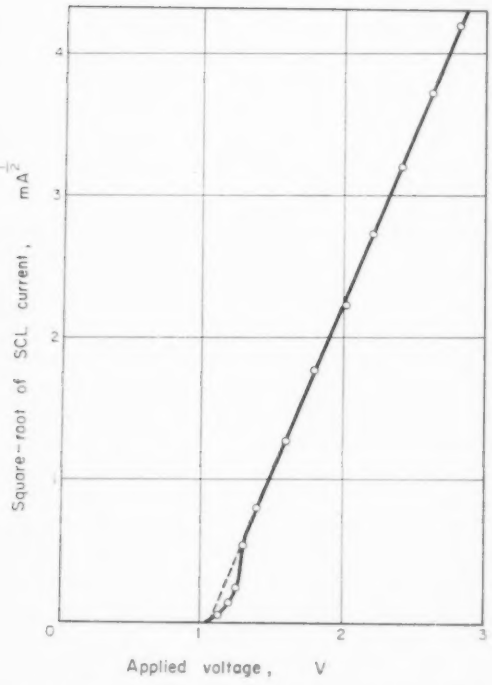


FIG. 3. Dependence of space-charge-limited steady current upon applied voltage for cadmium sulphide dielectric diode.

B. *High-frequency measurements*

Measurements of the high-frequency incremental admittance and of transit-time effects were made with the General Radio type 1062 B admittance meter over the frequency range from 50 to 900 Mc/s. At such high frequencies ordinary components have dimensions which are appreciable compared with the electromagnetic wavelength. However, the active region of the dielectric diode is sufficiently small that a lumped-element

definition is applicable. The diode was mounted at the end of a rigid coaxial line so that the cartridge containing the crystal formed an extension of the inner conductor and the mount holding the cartridge formed an extension of the outer conductor. A dummy diode was made which had exactly the same configuration as the real diode but did not contain the crystal. The basis of the measurement was to compare the admittance of the real diode with that of the dummy one. This was done by placing the dummy diode in the mount and adjusting the length of the rigid coaxial line with the admittance meter balanced at zero conductance and zero susceptance until a minimum output signal was obtained from the meter. This ensured that the coaxial line plus the leads in the dummy diode seen by the meter was an integral multiple of a half-wavelength long. The dummy diode was then replaced by the real diode and the conductance and susceptance arms of the meter adjusted until a minimum output was again obtained. The readings obtained in this way represented the admittance of the active region of the diode with the effects of anode and cathode leads in the cartridge eliminated.

In order to operate the meter with a current and voltage bias on the diode, it was necessary to insert coupling capacitors into the standard conductance and susceptance arms to isolate these from the steady-voltage source. This did not affect the standard conductance arm but a re-calibration of the short-circuit susceptance stud was necessary, since introduction of the coupling capacitor increased its length.

(i) *Measurement of incremental admittance.* We have seen that at low and moderate frequencies for which transit time effects are not significant the incremental equivalent circuit of the diode may be represented as a conductance in parallel with a capacitance. From equation (31) we see that, for this equivalent circuit, $rC = T_a/3$; this shows that the capacitive current will be comparable with the conductive current only at very high frequencies for which, in any event, transit-time effects will be significant. Evidently then, in order to measure these incremental parameters and confirm this equivalent circuit, it is necessary to use a frequency sufficiently high that the capacitive current is measurable and separable from the conductive current but not so high that transit time effects are observable. It was found that the sensitivity of the admittance meter was sufficient for this type of measurement to be made at the low-frequency end of its working range.

Measurements of incremental admittance under reverse-bias conditions were made at 50 Mc/s on a number of diodes; as expected the reverse conductance and capacitance remained constant. However, the reverse-bias conductance was finite, although small, instead of zero; this was attributed to dielectric losses in the crystal and cartridge. Measurements under forward-bias conditions showed that the conductance began to increase and the capacitance to decrease as the applied steady voltage was increased above the threshold V_t. This is to be expected for the inner potential difference across the crystal becomes positive as the applied voltage increases above V_t; space-charge enters the crystal and current occurs.

For the diode whose steady-current characteristic is shown in Fig. 3, an incremental resistance of 83 Ω was measured at a forward bias of 2 V. This compares well with the value of 90 Ω derived from the steady-current slope resistance of the diode.

The dependence of incremental capacitance upon bias voltage is shown for this diode in Fig. 4. In order to interpret these observations it will be recalled that the threshold voltage measures the point at which the inner potential difference across the diode is zero. Thus the measurements confirm that when the inner potential difference is reverse biased the capacitance is constant and that when it is forward biased the capacitance decreases to a somewhat lower value. The decrease is rather less than anticipated but this is to be expected, for there will inevitably be a certain amount of stray capacitance associated with the physical construction of the diode. In particular, if there exist regions of the cathode contact which have been insufficiently formed and are unable to inject electrons, these will add to the geometrical electrode capacitance but will not contribute to the space-charge-incremental capacitance. Such areas were observed by electroplating the anode contact on to the crystal, for plating was only deposited opposite the electron-injecting areas. In this way a map was formed of the effective regions of the cathode; it was found that these

might typically cover only part of the area in contact with the crystal. The areas of the cathode which were shown in this way to be non-injecting evidently added stray capacitance to the device.

These experimental observations confirm the essential features of the theoretical analysis for low and moderate frequencies at which transit time effects are not significant.

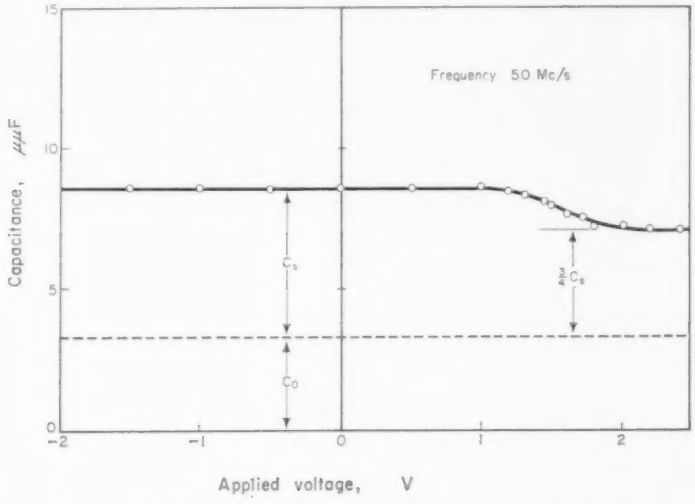


FIG. 4. Variation of low-frequency incremental input capacitance upon applied voltage (measured at 50 Mc/s).

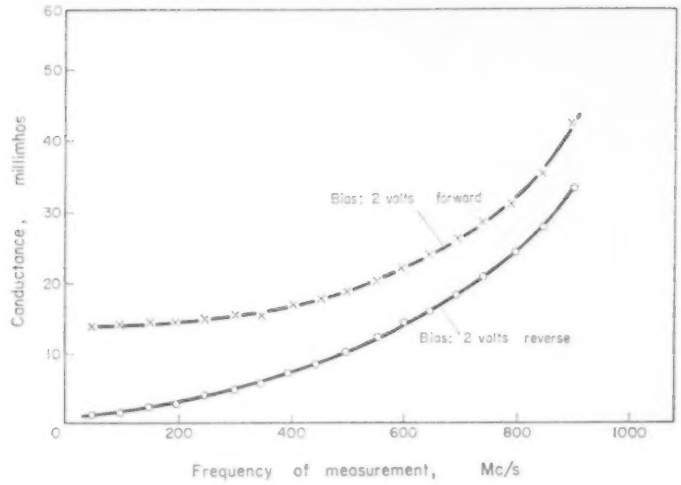
(ii) *Measurement of transit-time effects.* Transit-time effects were studied by measuring the incremental admittance of the diode over the whole frequency range of the admittance meter from 50 to 900 Mc/s.

The incremental admittance of the diode was first measured under reverse-bias conditions. Since the diode did not conduct with reverse-bias volts applied, the incremental conductance measured represented dielectric losses in the crystal and cartridge. These increased as the measuring frequency increased but at a given frequency did not change if the reverse bias voltage was varied. The incremental capacitance measured represented the sum of electrode capacitance and stray capacitance and remained constant as expected over most of the frequency range, although a slight increase was observed at the higher frequencies.

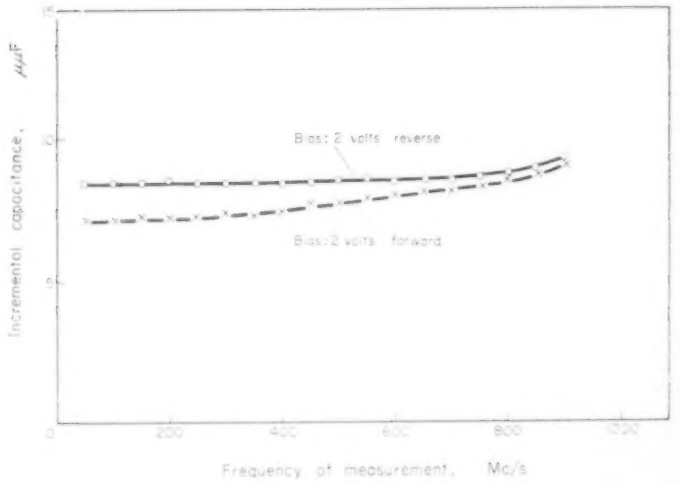
The incremental admittance was next measured with a constant steady forward-bias voltage of 2 V applied. Under these conditions the incremental conductance measured represented the sum of space-charge conduction and dielectric losses. The dependence of incremental conductance upon frequency is illustrated in Fig. 5(a); evidently the forward-bias space-charge conduction is the difference between these two curves. The incremental input capacitance measured represented the sum of space-charge capacitance and stray capacitance and increased slowly as the measuring frequency was increased. The dependence of input capacitance upon frequency is illustrated in Fig. 5(b); evidently the measurements of reverse-bias capacitance provide a normalization scale to separate the frequency dependence of the space-charge capacitance from any frequency dependence of capacitance that may exist, for example, because of dielectric dispersion in the crystal.

4. DISCUSSION

In the preceding Section we have described the experimental techniques used for the measurement



(a)



(b)

FIG. 5. Illustrating the variation with frequency of incremental conductance and of incremental input capacitance under forward- and reverse-bias conditions.

of the very high-frequency characteristics of the diodes. In this Section we compare the experimental observations with the theoretical calculations.

A. *Correction for stray capacitance*

It has been remarked previously that there is inevitably a certain amount of stray capacitance associated with the structure of the diode. This

can be estimated if we accept the theoretical result that the forward-biased incremental capacitance is three-quarters of the geometrical capacitance. This seems justified, because the form of the experimental results, showing in particular that the forward-biased space-charge capacitance is smaller than the reverse-biased geometrical capacitance, indicates the correctness of the theoretical calculations for low and moderate frequencies, and there seems no reason to doubt the actual numerical ratio indicated.

If we denote by C_s and C_0, respectively, the geometrical capacitances of the injecting and non-injecting areas of the cathode, we have from the reverse-bias portion of Fig. 4 that

$$C_s+C_0 = 8{\cdot}5\ \mu\mu\text{F}$$

Using the forward-bias portion of this Figure we have

$$\frac{3C_s}{4}+C_0 = 7{\cdot}2\ \mu\mu\text{F}$$

Thus

$$C_s = 5{\cdot}2\ \mu\mu\text{F}$$
$$C_0 = 3{\cdot}3\ \mu\mu\text{F}$$

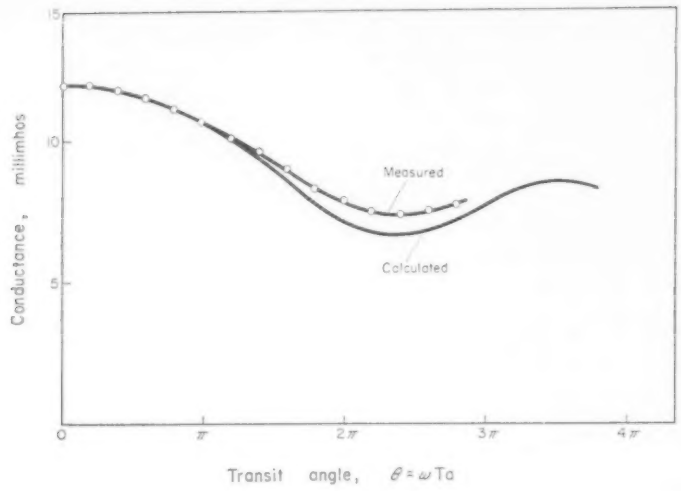
This indicates that in this particular diode a fraction, $C_s/(C_s+C_0) = 0{\cdot}6$, of the cathode area was actually effective in injecting electrons and that the remainder added stray capacitance. This division of the measured capacitance between stray capacitances and diode capacitance has been indicated in Fig. 4. Evidently it would be desirable to measure the stray capacitance separately; however, there seems no obvious way of doing this.

B. *Comparison of theory with experiment*

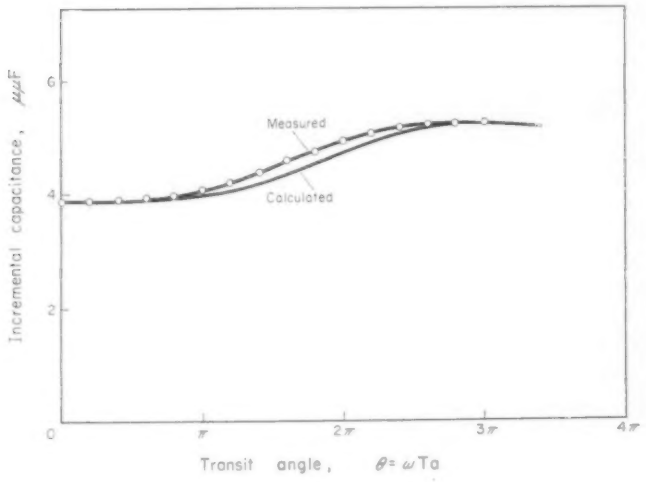
The most convenient way to compare the theoretical calculations with the experimental results is to express both in terms of the transit angle θ. This has already been done for the former and may readily be done for the latter.

We have that $\theta = \omega T_a$ and that $T_a = 3rC_s$. For the diode for which the experimental results are given at a forward bias of 2 V, the incremental resistance is 83 Ω and the effective geometrical capacitance is $5{\cdot}2\ \mu\mu$F; these figures give an electron transit time of $1{\cdot}5\ \text{m}\mu$ sec. Thus the frequency at which the period equals the transit time is 670 Mc/s and lies within the frequency range used. Using that the frequency of 670 Mc/s corresponds to a transit angle of 2π, the experimental results may be re-plotted in terms of the transit angle. This has been done for the incremental conductance in Fig. 6(a) and for the incremental capacitance in Fig. 6(b); the stray capacitance of $3{\cdot}3\ \mu\mu$F has been subtracted from the measured capacitance in order to plot the latter curve. The theoretical curves are also given on these diagrams; to facilitate comparison these are normalized to the experimental values at low frequencies for which transit-time effects are not significant. The agreement between the two sets of curves is very satisfactory and demonstrates that the physical model adopted and the theoretical analysis presented may be used with confidence to describe the small-signal behaviour of the SCL dielectric diode at all frequencies.

It is relevant at this point to recall that the experimental measurements presented refer to a dielectric diode constructed from a cadmium sulphide crystal. This material is readily grown in the form of thin crystal plates, is economic and is suitable for the demonstration of basic principles. This is particularly so for the present purpose, because the relatively small mobilities of charge carriers in this material increases the transit time of electrons across the device to such an extent that measurements of transit-time effects can be made in the sub-kilomegacycle frequency range. For device purposes, however, it is unsuitable; materials such as gallium arsenide and aluminium antimonide with their much higher mobilities for charge carriers providing higher operating frequencies and higher forward conductances would seem far more attractive for the development of SCL solid-state devices.[1] In this connexion it is interesting to observe that for a dielectric diode made from gallium arsenide, in which the electrode spacing is $10\,\mu$, and operated at an applied inner potential of 1 V, the frequency at which the transit angle becomes 2π is 7 kMc/s; if the electrode spacing is reduced to $3\,\mu$ the corresponding frequency is 70 kMc/s and if the spacing is reduced to $1\,\mu$ the frequency is 700 kMc/s. This reduction in electrode spacing can be made without unduly sacrificing reverse breakdown strength;[1] it is apparent therefore



(a)



(b)

FIG. 6. Comparison of theoretical calculations and experimental measurements of incremental conductance and incremental input capacitance.

that the SCL dielectric diode is a potentially useful device for very high-frequency applications.

It is interesting to observe that transit-time effects do not introduce a negative resistance at very high frequencies. This does occur in the vacuum diode because the momentum of the injected space charge prevents it from responding immediately to the time variations of the applied electric field. In the dielectric diode, however, the momentum of the injected space-charge is destroyed by collisions of the individual electrons with the crystal lattice and these therefore respond

almost instantaneously to the time variations of the applied electric field. Accordingly, in the solid-state case the space charge always moves in the direction of the electric-field intensity; at very high frequencies, however, the space charge is not able to cross the electrode space within the half-period of the applied frequency and consequently the conductance does slowly decrease as the frequency increases.

5. CONCLUSIONS

The small-signal theory presented shows that the incremental equivalent circuit may be represented as a conductance in parallel with a capacitance. This equivalent circuit is valid for all frequencies, but, at frequencies sufficiently high that transit time effects begin to appear, the parameters become frequency dependent. For these very high frequencies, expressions are derived to enable the conductance and susceptance to be calculated in terms of transit angle.

The experimental measurements made with coaxial cadmium sulphide dielectric diodes compare very satisfactorily with the calculated parameters and show that the theoretical approach adopted and the physical model used for the mechanisms of SCL current in solids are essentially correct.

The theory and experiment presented here show that the SCL dielectric diode is capable of operation up to very high frequencies, particularly if constructed from high-mobility materials such as gallium arsenide or aluminium antimonide. Since the diode should also possess a low inherent noise figure and be relatively insensitive to temperature changes it is a potentially useful device.

Acknowledgements—We would like to thank Professor D. G. Tucker of the Electrical Engineering Department for the provision of general experimental facilities for this research, and the Department of Scientific and Industrial Research for a Research Grant to purchase the admittance meter used. We would also like to thank the Electrical Engineering Department of the Aeronautical College, Cranfield, for providing facilities for preliminary high-frequency measurements. One of us (J. S.) would like to thank the University of Birmingham for a University Scholarship during the period of this research.

REFERENCES

1. G. T. WRIGHT, *I.R.E. Solid-State Circuits Conference, Philadelphia*, 1961. Paper 1:4.
2. W. E. BENHAM, *Phil. Mag.* **5**, 641 (1928).
3. F. B. LLEWELLYN, *Bell Syst. Tech. J.* **14**, 632 (1935).
4. G. T. WRIGHT, *Solid-State Electron.* **2**, 165 (1961).
5. G. T. WRIGHT, *Proc. Instn. Elect. Engrs.* **B106**, 915 (1959).

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 304–308. Printed in Great Britain

INFLUENCE OF WAFER THICKNESS AND CARRIER RECOMBINATION ON THE CUTOFF FREQUENCY OF ALLOY-JUNCTION TRANSISTORS

S. AMER

Standard Telephones and Cables Limited, Transistor Division, Footscray, Kent

(*Received* 16 *June* 1961; *in revised form* 7 *August* 1961)

Abstract—The influence of wafer thickness and carrier recombination on the α-cutoff frequency of the "intrinsic" transistor is studied. First, MOYJES' solution of the diffusion of minority carriers in a symmetrical junction transistor is used to find the d.c. transport factor. This is then converted to the a.c. case and finally an implicit equation is found for the cutoff frequency as a function of wafer thickness θ and surface recombination velocity S. It is shown that, for $S < 3000$ cm/sec, the influence of surface recombination is negligible.

At very narrow basewidths the current density is very high at the emitter edge and this decreases the diffusion length. More important seems to be the fact that the carriers' path (from slanted edge of emitter to slanted edge of collector) is longer than the geometrical basewidth. A reduction in wafer thickness decreases this effect and thus increases the cutoff frequency, in spite of increasing the current density at the emitter edge. The most important equations have been represented graphically for the particular case of the STC TK20 transistor.

Résumé—L'influence de l'épaisseur de la tranche et de la recombinaison des porteurs sur la fréquence de coupure Alpha du transistor intrinsèque est étudiée. Tout d'abord, la solution de MOYJES pour la diffusion de porteurs minoritaires dans un transistor à jonction symétrique est employée pour trouver le facteur de transport à courant continu. On convertit ensuite cette solution pour le cas du courant alternatif et finalement une equation absolue est trouvée pour la fréquence de coupure en fonction de l'épaisseur de tranche θ et de la vitesse de recombinaison de surface S. Il est démontré que pour $S < 3000$ cm/sec, l'influence de recombinaison de surface est négligeable.

A des largeurs de base très limitées, la densité du courant est très élevée à la borne de l'émetteur et ceci réduit la longueur de diffusion. Ce qui paraît être beaucoup plus important est le fait que le parcours des porteurs (du bord incliné de l'émetteur au bord incliné du collecteur) est plus long que la largeur géometrique de la base. Cet effet est limité si l'épaisseur de la tranche est réduite, ce qui accroît aussi la fréquence de coupure, malgré l'augmentation de la densité de courant au bord de l'émetteur. Les équations les plus importantes sont représentées graphiquement pour le cas particulier du transistor STC TK20.

Zusammenfassung—Der Einfluss der Schichtdicke und der Träger-Rekombination auf die α-Grenzfrequenz des "eigentlichen" Transistors wird untersucht. Unter Verwendung von MOYJES' Lösung für die Diffusion von Minoritätsträgern in einem symmetrischen Übergangs-Transistor wird der Transportfaktor für Gleichstrom bestimmt. Er wird dann auf den Fall von Wechselstrom umgerechnet, und es ergibt sich schliesslich eine Gleichung, die die Grenzfrequenz als Funktion der Schichtdicke θ und der Geschwindigkeit der Rekombination an der Oberfläche, S, enthält. Für $S < 3000$ cm/sec ist der Einfluss dieser Rekombination vernachlässigbar.

Bei sehr geringer Basisdicke ist die Stromdichte an der Emitterkante sehr hoch, und dies verringert die Diffusionslänge. Noch wichtiger erscheint die Tatsache, dass der vom Träger durchlaufene Weg (von der schrägen Emitterkante bis zur schrägen Kollektorkante) länger ist als die geometrische Basisdicke. Eine Reduzierung der Schichtdicke vermindert diesen Effekt und erhöht damit die Grenzfrequenz, trotz der Zunahme der Stromdichte an der Emitterkante. Die wichtigsten Gleichungen wurden für den Sonderfall des STC TK20-Transistors graphisch dargestellt.

1. INTRODUCTION

THE cutoff frequency of the "intrinsic" transistor is limited by carrier dispersion during the transit time across the base. Previous studies (see for instance Refs. 1 and 2) have considered only carrier dispersion inside the base region. This will be extended to the more general case when the carriers' path may be longer than the geometrical basewidth and also to include carrier losses by surface and volume recombination outside the base region. Emitter efficiency is assumed to be very near unity and thus independent of frequency.[3] At very high frequencies cutoff will be limited by junction capacitances and base resistance,[4,5] but, for transistors with a total f_{hb} not very high, the limiting factor may be the "intrinsic" f_{hb} and thus the mechanism proposed here becomes important.

2. DIFFUSION OF MINORITY CARRIERS (D.C. CASE)

The diffusion of minority carriers in the d.c. case for a symmetrical transistor has been solved by MOYJES[6]. MOYJES has used the geometry of Fig. 1 and his results are

$$I_s = I_E \frac{\pi SW\theta R}{A_E D\left(\frac{S\theta}{2D}+\frac{\theta^2}{4L^2}\right)^{1/2}} \frac{K_1\left[\frac{2R}{\theta}\left(\frac{S\theta}{2D}+\frac{\theta^2}{4L^2}\right)^{1/2}\right]}{K_0\left[\frac{2R}{\theta}\left(\frac{S\theta}{2D}+\frac{\theta^2}{4L^2}\right)^{1/2}\right]} \tag{1}$$

and

$$I_V = I_E\left[\frac{W^2}{2L^2}+\frac{\pi\theta^2 WR}{2L^2A_E\left(\frac{S\theta}{2D}+\frac{\theta^2}{4L^2}\right)^{1/2}} \frac{K_1\left[\frac{2R}{\theta}\left(\frac{S\theta}{2D}+\frac{\theta^2}{4L^2}\right)^{1/2}\right]}{K_0\left[\frac{2R}{\theta}\left(\frac{S\theta}{2D}+\frac{\theta^2}{4L^2}\right)^{1/2}\right]}\right] \tag{2}$$

where

I_s = surface recombination current
I_E = emitter current
S = surface recombination velocity
W = basewidth
θ = wafer thickness
R = emitter radius
A_E = emitter area
D = diffusion constant of minority carriers in base
L = diffusion length of minority carriers in base
K_1 and K_0 = Bessel functions of imaginary argument
I_V = volume recombination current (both inside and outside the base region)

The argument of the Bessel functions in (1) and (2) is fairly large for all the normal values of the parameters found in transistors. This allows a great simplification of formulae (1) and (2). The Bessel functions can be substituted by their asymptotic expansions and their ratio is practically unity.

Thus, since the only losses of minority carriers are due to surface and volume recombination, the transport factor is given by

$$\beta = 1-\frac{W^2}{2L^2}-\frac{\pi\theta^2WR}{2L^2A_E[(S\theta/2D)+(\theta^2/4L^2)]^{1/2}} - \frac{\pi SW\theta R}{A_E D[(S\theta/2D)+(\theta^2/4L^2)]^{1/2}} \tag{3}$$

For a useful transistor the transport factor is very close to unity, and (3) can be written as

$$\frac{1}{\beta} = 1+\frac{W^2}{2L^2}+\frac{\pi WR}{A_E}\left(\frac{\theta^2}{L^2}+\frac{2S\theta}{D}\right)^{1/2} \tag{4}$$

3. TRANSPORT FACTOR IN THE A.C. STEADY STATE CASE

As pointed out by STEELE[1], the only difference between the d.c. and the a.c. steady-state diffusion equations is that where $1/L^2$ appears in the first, $(1/L^2)\ (1+j\omega\tau)$ appears in the second; τ is the

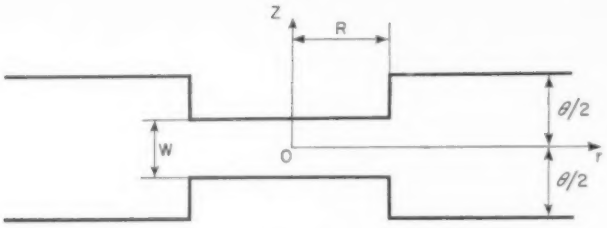


FIG. 1.

lifetime of minority carriers in the base. Thus we can write for the a.c. steady-state case

$$\frac{1}{\beta} = 1+\frac{W^2}{2L^2}(1+j\omega\tau) + \frac{\pi WR}{A_E}\left(\frac{\theta^2}{L^2}(1+j\omega\tau)+\frac{2S\theta}{D}\right)^{1/2} \tag{5}$$

For fairly high frequencies ($\omega > 5\times 10^6$) and not very low lifetimes ($\tau > 10\,\mu$sec) (5) can be simplified to

$$\frac{1}{\beta} = 1+j\frac{W^2\omega}{2D}+\frac{\pi WR}{A_E}\left(j\frac{\theta^2\omega}{D}+\frac{2S\theta}{D}\right)^{1/2} \tag{6}$$

4. CALCULATION OF THE CUTOFF FREQUENCY

In the common base response, the cutoff frequency ω_α, defined by a fall in power gain of 3 dB, corresponds to $\beta = 1/\sqrt{2}$ and thus we can write from (6)

$$2 = \left[1+\frac{\pi WR}{A_E}\left\{\frac{\theta}{2D}[2S+(4S^2+\theta^2\omega_\alpha{}^2)^{1/2}]\right\}^{1/2}\right]^2 + \left[\frac{W^2\omega_\alpha}{2D}+\frac{\pi WR}{A_E}\left\{\frac{\theta}{2D}[(4S^2+\theta^2\omega_\alpha{}^2)^{1/2}-2S]\right\}^{1/2}\right]^2 \tag{7}$$

It is obvious from (7) that if surface recombination and volume recombination outside the base region are neglected, and only the diffusion across the base region is considered, we obtain the normally quoted formula

$$\omega_\alpha = 2D/W^2 \tag{8}$$

For cutoff frequencies higher than $f_{hb} = 2$ Mc/s (i.e. $\omega_\alpha = 4\pi 10^6$) and surface recombination velocities smaller than $S = 3000$ cm/sec, i.e. for ordinary conditions, the terms containing S in (7) are negligible and we can reduce (7) to

$$2 = \left[1+\frac{\pi WR\theta}{A_E}\left(\frac{\pi f_{hb}}{D}\right)^{1/2}\right]^2 + \left[\frac{\pi W^2 f_{hb}}{D}+\frac{\pi WR\theta}{A_E}\left(\frac{\pi f_{hb}}{D}\right)^{1/2}\right]^2 \tag{9}$$

This is a fourth-degree equation giving f_{hb} as a function of W and θ. It cannot in general be simplified without ignoring significant terms.

5. CUTOFF FREQUENCY AT VERY NARROW BASEWIDTHS

At very narrow basewidths we can find f_{hb} from (9) by finding

$$\lim_{W\to 0} f_{hb}$$

But first we have to find a suitable expression for A_E, since at very narrow basewidths only a narrow ring around the emitter edge will be actually working as an emitter. This is done in the Appendix and the result is

$$A_E = 2\pi CWR\,I_1(R/CW)/I_0(R/CW) \tag{10}$$

where C is a constant, characteristic of each type of transistor, that has to be determined experimentally and I_0 and I_1 are Bessel functions of

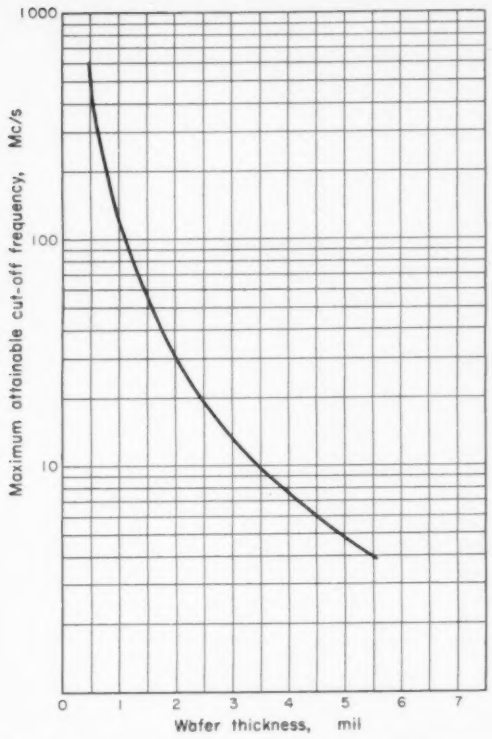


FIG. 2. TK 20—Maximum attainable cutoff frequency (for zero basewidth) as a function of wafer thickness.

imaginary argument. Substituting (10) in (9) we find

$$2=\left[1+\frac{\theta I_0(R/CW)}{2CI_1(R/CW)}\left(\frac{\pi f_{hb}}{D}\right)^{1/2}\right]^2 + \left[\frac{\pi W^2 f_{hb}}{D}+\frac{\theta I_0(R/CW)}{2CI_1(R/CW)}\left(\frac{\pi f_{hb}}{D}\right)^{1/2}\right]^2 \tag{11}$$

and from (11)

$$\lim_{W\to 0} f_{hb} = 0{\cdot}17\frac{DC^2}{\theta^2} \tag{12}$$

From (12) we can evaluate C if

$$\lim_{W\to 0} f_{hb}$$

is known experimentally. For instance, for the STC TK20 transistor with $\theta = 2$ mils we know $f_{hbm} = 30$ Mc/s and thus we obtain

$$C = 10{\cdot}17 \tag{13}$$

As can be seen from (10), the physical meaning of C is the width (measured in units of W) of the effective emitter ring at very low values of the basewidth. The value given by (13) thus appears to be plausible.

Taking again as a numerical example the TK20, where $R = 10$ mils, we can write equation (11) as

$$2=\left[1+0{\cdot}033 f_{hb}^{1/2}\frac{I_0(0{\cdot}98/W)}{I_1(0{\cdot}98/W)}\right]^2 + \left[0{\cdot}46W^2 f_{hb}+0{\cdot}033 f_{hb}^{1/2}\frac{I_0(0{\cdot}98/W)}{I_1(0{\cdot}98/W)}\right]^2 \tag{14}$$

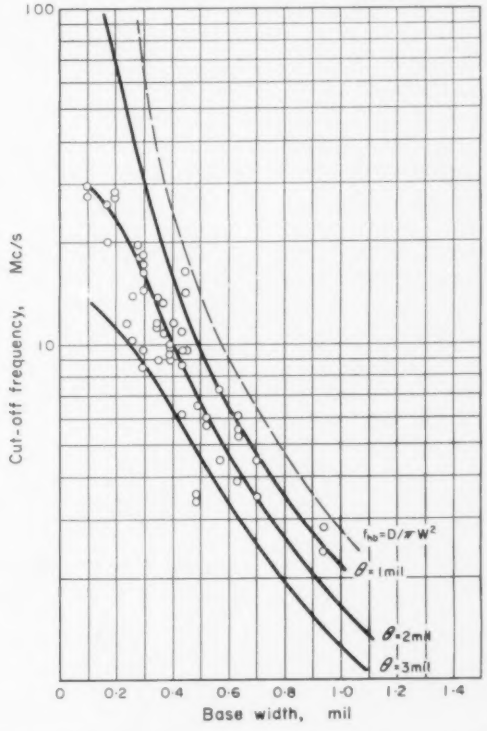


FIG. 3. TK 20's—Cutoff frequency against basewidth for different values of the wafer thickness; depletion layer of 0·1 mil width has been allowed for comparison with experimental points marked +.

In Fig. 2 the parabola (12) has been plotted for the particular case of the TK20's, i.e. with the value of C given by (13).

Comparison with experiment is made in Fig. 3 where (14) has been plotted for three values of wafer thickness.

6. CONCLUSIONS

It has been shown that surface and volume recombination play a very small part in limiting the cutoff frequency of alloy transistors, such as the TK20.

For narrow basewidths, the current density near the emitter edge is very high and this leads to smaller diffusion lengths and a decrease in cutoff frequency. This has been taken into account by introducing an effective emitter area.

More important seems to be the fact that the actual path of carriers from emitter to collector is longer than the geometrical basewidth, when the latter is very narrow. This is because most carriers are injected at the slanted edge of the emitter and collected at the slanted edge of the collector. A reduction in wafer thickness will decrease this effect and thus raise the cutoff frequency, in spite of increasing the carrier density near the emitter edge.

Acknowledgements—Acknowledgement is made to Standard Telephones and Cables Limited for permission to publish this paper; also the help of Dr. R. I. B. COOPER and Mr. T. J. ROWE, who pointed out MOYJES' work on diffusion and the previous work carried out at Standard Telephones and Cables Limited on this subject. Miss W. A. GREENFIELD carried out the measurements recorded in Fig. 3.

REFERENCES

1. E. L. STEELE, *Proc. I.R.E.* **40,** 1427 (1952).
2. W. SHOCKLEY, M. SPARKS and G. K. TEAL, *Phys. Rev.* **83,** 161 (1951).
3. R. L. PRITCHARD, *Proc. I.R.E.* **46,** 1137 (1958).
4. F. J. HYDE, *Proc. Instn. Elect. Engrs.* **B106,** 391 (1959).
5. J. LINDMAYER and C. WRIGLEY, *Solid-State Electron.* **2,** 247 (1961).
6. B. Y. MOYJES, *Soviet Phys. Tech. Phys.* **28,** 2402 (1958).

APPENDIX

Effective Emitter Area at very narrow Basewidths

At very narrow basewidths the self-biasing effect due to the base current is so great that the current density is very high and diffusion constants and lengths take quite different values from those at low-level injection.

The diffusion length will thus depend on basewidths and, as a first approximation, we can assume a linear relationship

$$L = CW \tag{A.1}$$

The radial diffusion equation of carriers will obey the following equation

$$\frac{d^2N}{dr^2} + \frac{1}{r}\frac{dN}{dr} = \frac{N}{L^2} = \frac{N}{C^2W^2} \tag{A.2}$$

the solution of which is

$$N = N_0 I_0(r/CW) \tag{A.3}$$

where I_0 is Bessel's function of imaginary argument and order zero, and N_0 the injected carrier density at the emitter centre.

If the effective emitter area A_{ett}, is defined as that area where the product A_{eff} times the current density at the emitter edge is equal to the total emitter current, then we have

$$A_{\text{eff}} = 2\pi CWR\frac{I_1(R/CW)}{I_0(R/CW)} \tag{A.4}$$

For small W, i.e. $R/CW \gg 1$, we find

$$A_{\text{eff}} = 2\pi CWR \tag{A.5}$$

and for large W, i.e. $R/CW \ll 1$, we find

$$A_{\text{eff}} = \pi R^2 \tag{A.6}$$

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 309–322. Printed in Great Britain

NOTES

Nomograph technique for doping determination in germanium and silicon crystal growing

(*Received* 10 *August* 1961)

I. *Introduction*

IN CRYSTAL growing, zone leveling techniques are employed to produce single crystals of a desired, uniform resistivity.[1] In order to determine the impurity requirements for doping the liquid zone, a number of factors such as atomic weight of the element, segregation coefficient, and drift mobility need to be considered. These require individual calculations for each particular doping run based upon the element used and the resistivity desired. This paper derives a generalized, doping equation which can be utilized in nomograph form to determine any impurity requirement.

II. *Theory*

The zone leveling technique is most effective when the segregation coefficient for the contained impurity is less than 0·1. Segregation coefficient, K, is defined as the ratio of the solubility of impurity in the solid, C_s, to the solubility of the impurity in the liquid, C_l: $(K = C_s/C_l)$. The concentration, C_s, in the solid after moving the zone a finite distance is related to the resistivity as $C_s = 1/\rho e\mu$. This leads to an expression for C_l as follows:

$$C_l = \frac{C_s}{K} = \frac{(1/\rho e\mu)}{K} \tag{1}$$

where e = electron charge $1{\cdot}6 \times 10^{-19}$ (c), μ = electron or hole drift mobility (cm^2/V sec), and ρ = resistivity (Ω cm).

Constant mobility condition. The approximate value of drift mobility for resistivities ranging between 0·1 to 20 Ω cm is 1800 cm^2/V sec for P-type, and 3600 cm^2/V sec for N-type germanium. These values of mobility are assumed to be constant over this resistivity range. The weight of impurity atoms, w_o, necessary to dope a unit volume (cm^3) of the liquid zone to a particular impurity concentration, C_l, is:

$$w_o = C_l(n/N) \tag{2}$$

where w_o = gram impurity per unit volume of the liquid zone, N = Avogadro's Number $(6{\cdot}02 \times 10^{23})$ and n = atomic weight of the impurity element. Combining equations (1) and (2), we obtain a generalized doping equation for the constant mobility condition:

$$w_o\rho = [(N/n)e\mu K]^{-1} \tag{3}$$

Substituting the numerical values for e and μ, equation (3) reduces to:

$$w_o\rho = 2{\cdot}88 \times 10^{-9}(n/K)(\text{for } N\text{-type}) \tag{4}$$

$$w_o\rho = 5{\cdot}77 \times 10^{-9}(n/K)(\text{for } P\text{-type}) \tag{5}$$

For any individual impurity element the expression n/K is constant. Therefore, equation (3) can be written as:

$$\rho w_o = R\,(\text{constant}) \tag{6}$$

This equation states that, for any impurity element, the product of number of impurity grams introduced (per unit volume of the liquid zone) and the resistivity obtained is a constant. Table 1 lists the R values for several impurity elements in germanium.

Equation (6) may be plotted for any impurity element, once the value of constant R for the element is determined. Fig. 1 shows a log–log plot of the equation for indium $(R = 6{\cdot}02 \times 10^{-4})$. From this graph, the weight of the indium required to obtain any resistivity can readily be determined. An approach of this type has the following restrictions: (1) the drift mobility is considered constant, and (2) a separate graph is required for each doping element.

Non-constant mobility condition. In order to derive a general equation to be used for any element and to consider the effect of mobility as a function of resistivity, the following procedures

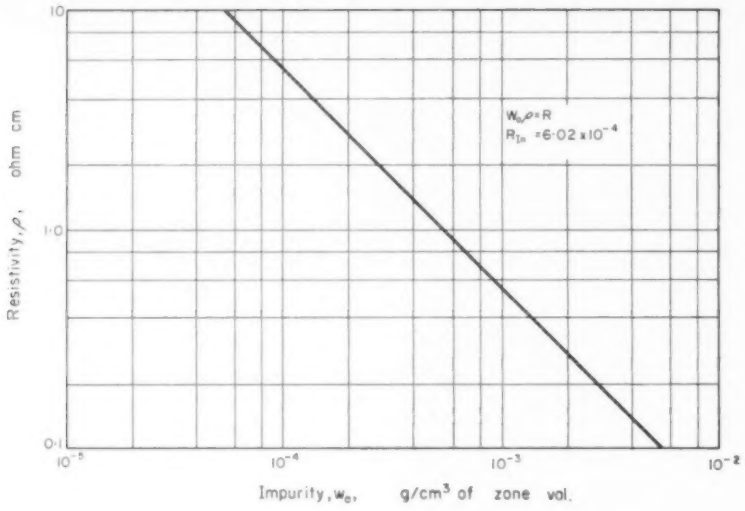


FIG. 1. Resistivity vs. weight of indium as an impurity element in germanium [$R_{In} = 6{\cdot}02 \times 10^{-4}$].

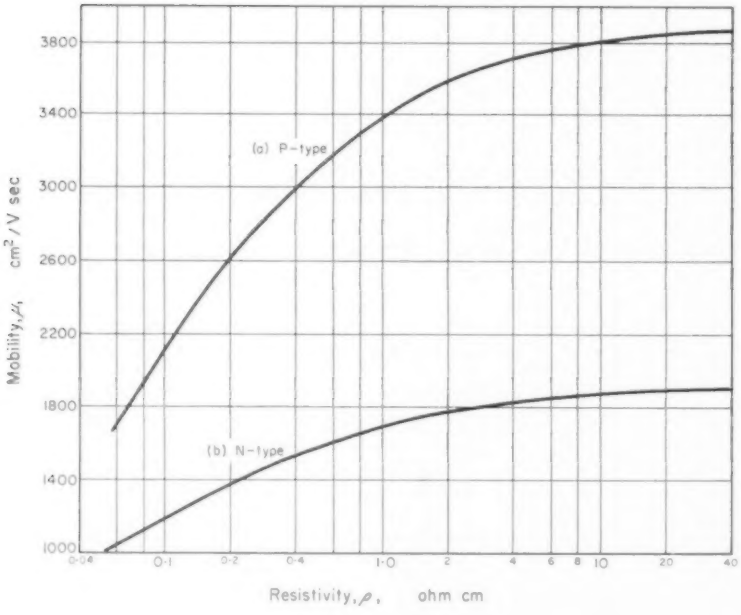


FIG. 2. (a) Electron drift mobility vs. resistivity of *P*-type germanium at 300° K. (b) Hole drift mobility vs. resistivity of *N*-type germanium at 300 °K.

Table 1.

R_o values for impurities in germanium for constant mobility condition

Element	Atomic wt. n	Seg. coeff. K[2]	Constant R
B	10·82	~ 20	$3{\cdot}12\times10^{-9}$
Al	26·98	1×10^{-1}	$1{\cdot}56\times10^{-6}$
Ga	69·72	1×10^{-1}	$4{\cdot}02\times10^{-6}$
In	114·82	$1{\cdot}1\times10^{-3}$	$6{\cdot}02\times10^{-4}$
Tl	204·39	4×10^{-5}	$2{\cdot}95\times10^{-2}$
P	30·97	$1{\cdot}2\times10^{-1}$	$7{\cdot}43\times10^{-7}$
As	74·91	4×10^{-2}	$5{\cdot}39\times10^{-6}$
Sb	121·76	$3{\cdot}0\times10^{-3}$	$1{\cdot}17\times10^{-4}$
Bi	209·0	4×10^{-5}	$1{\cdot}51\times10^{-2}$

may be used. Fig. 2 shows the theoretical curves for hole and electron drift mobility versus resistivity for germanium at 300 °K.[3] Fig. 3 is a plot of equation $C_s = 1/\rho e\mu$, using the mobility data from Fig. 2. Since the plot is linear for resistivities between 0·1 and 20 Ω cm., the resistivity impurity concentration relationships can be expressed by the following equations:

$$N_d = 1{\cdot}73\times10^{15}/\rho \text{ (for } N\text{-type)} \quad (7)$$

$$N_a = 3{\cdot}69\times10^{15}/\rho \text{ (for } P\text{-type)} \quad (8)$$

where N_d and N_a are total concentration of donor and acceptor impurity atoms, (atmos/cm^3). Substituting the values of N_d and N_a into equation (2) and solving for $w_o\rho$, we get:

$$w_o\rho = 2{\cdot}87\times10^{-9}(n/K) \text{ (for } N\text{-type)} \quad (9)$$

$$w_o\rho = 6{\cdot}12\times10^{-9}(n/K) \text{ (for } P\text{-type)} \quad (10)$$

The right hand side of equations (9) and (10) has a constant value for each impurity element. Therefore:

$$w_o\rho = R_o \text{ (constant)} \quad (11)$$

Since $w_o = w/v$, equation (11) can be written:

$$w\rho = R_o v \quad (12)$$

where w = total impurity required (g), and v = liquid zone volume (cm^3). A list of the R_o values to be used in equation (12) for several impurity elements are shown in Table 2 for germanium and Table 3 for silicon.

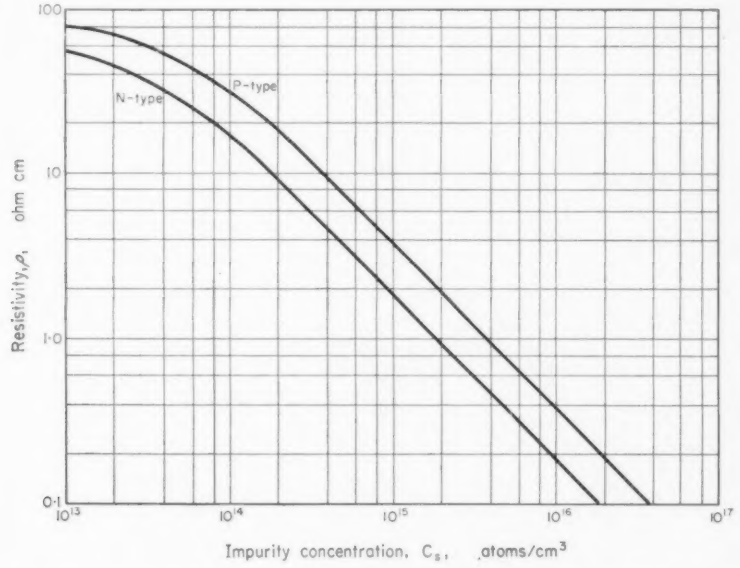


FIG. 3. Resistivity vs. absolute impurity concentration for germanium at 25 °C.

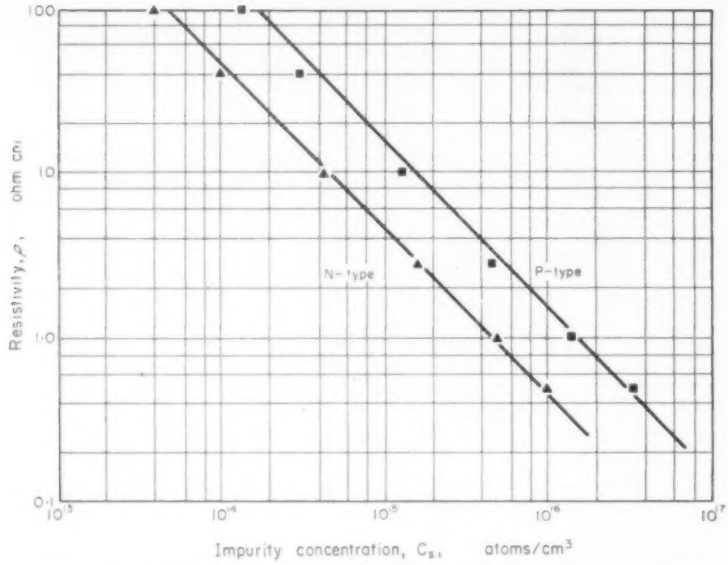


FIG. 4. Resistivity vs. absolute impurity concentration for silicon at 25 °C.

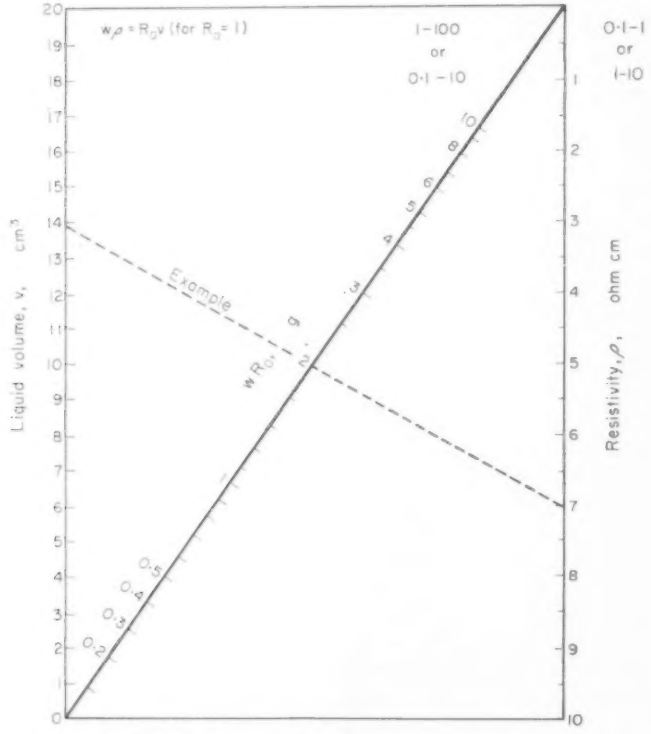


FIG. 5. Doping nomograph for liquid volumes from zero to 20 cm³.

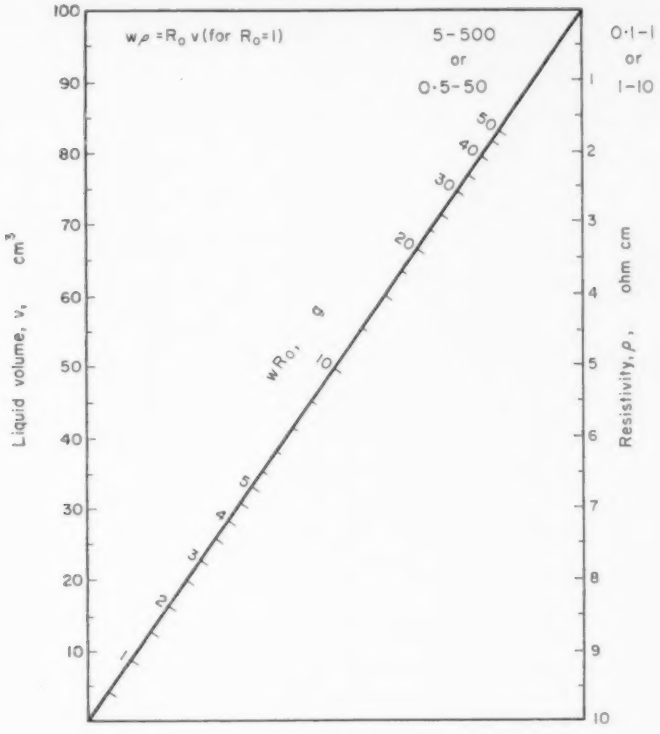


FIG. 6. Doping nomograph for liquid volumes from zero to 100 cm³.

Table 2.

R_0 values for impurities in germanium, for variable mobility condition

	Elements	R_0
P-type	B	$3{\cdot}31\times10^{-9}$
	Al	$1{\cdot}65\times10^{-6}$
	Ga	$4{\cdot}27\times10^{-6}$
	In	$6{\cdot}39\times10^{-4}$
	Tl	$3{\cdot}13\times10^{-2}$
N-type	P	$7{\cdot}41\times10^{-6}$
	As	$5{\cdot}38\times10^{-6}$
	Sb	$1{\cdot}16\times10^{-4}$
	Bi	$1{\cdot}50\times10^{-2}$

III. *Applications of nomograph*

Figure 5 illustrates a standard nomograph plot[4] of equation (12) for $R_0 = 1$, and relates the three interdependent variables i.e. resistivity, total weight of the impurity element, and volume of liquid zone. The number of grams of impurity required for doping is determined by drawing a straight line between the zone volume and the desired resistivity point. The intersection of this line with the wR_0 scale gives the number of grams of impurity required. It should be noted that this scale is calibrated for $R_0 = 1$. Therefore, the w value obtained from the scale must be multiplied by the proper R_0 value for the doping element being used. An example showing the use of this nomograph for indium is given in the Appendix.

The nomograph (Fig. 5) is constructed with the liquid volume scale restricted to a maximum of 20 cm³, which is most useful in horizontal crystal growing. If necessary, a similar nomograph with other volume scales may be constructed to conform with a particular crystal growing condition. Fig. 6 shows a nomograph constructed for volume and

Table 3.

R_o values for impurities in silicon, for variable mobility condition

Element	Atomic wt. n	Seg. coeff. $K^{(2)}$	Constant R_0
B	10·82	$8{\cdot}8\times10^{-1}$	$3{\cdot}9\times10^{-7}$
Al	26·98	$1{\cdot}6\times10^{-3}$	$4{\cdot}2\times10^{-4}$
Ga	69·72	$4{\cdot}0\times10^{-3}$	$4{\cdot}3\times10^{-4}$
In	114·82	$3{\cdot}0\times10^{-4}$	$9{\cdot}5\times10^{-3}$
P	30·97	$4{\cdot}0\times10^{-2}$	$5{\cdot}8\times10^{-6}$
As	74·91	$7{\cdot}0\times10^{-2}$	$8{\cdot}0\times10^{-6}$
Sb	121·76	$1{\cdot}8\times10^{-2}$	$5{\cdot}1\times10^{-5}$

resistivity ranges normally used in Czochralski crystal growing.

In most zone leveling and Czochralski crystal growing operations, the liquid zone volume can be determined quite accurately. Therefore, by using such a nomograph, the exact weight of a doping element for any resistivities between 0·1 and 20 Ω cm can be easily determined. The use of this nomograph has the following advantages: (1) it eliminates repeated doping calculations; (2) it requires only one principal nomograph and table of R_o values to determine any impurity doping requirement; (3) the table of R_o allows a comparison of relative weights of the various elements required to obtain any resistivity.

Acknowledgement—We wish to acknowledge the efforts of Mr. E. SHERWIN, who assisted in the preparation of this paper and the necessary experimental evaluations.

Raytheon Company
Semiconductor Division
Newton
Massachusetts

N. HEMMAT
A. L. MACDONALD

Appendix

Growth conditions

(a) Horizontal germanium crystal growing
(b) Liquid zone volume —14 cm³
(c) Desired resistivity—0·7 Ω cm
(d) Doping element—indium.

From Fig. 5, $w = 20$ and from Table 2, for In, $R_0 = 6{\cdot}39\times10^{-4}$. Therefore, the number of grams of In required is equal to $wR_0 = 20\times6{\cdot}39\times10^{-4} = 12{\cdot}78$ mg.

References

1. W. G. PFANN, *Zone Melting* p. 153. John Wiley, New York (1958).
2. W. C. DUNLAP, JR., *An Introduction to Semiconductors* p. 210. John Wiley, New York (1957).
3. M. B. PRINCE, *Phys. Rev.* **92**, 861 (1953).
4. A. S. LEVENS, *Nomography* (2nd Ed.). John Wiley, New York (1959).

Electrode geometries for which the transverse magnetoresistance is equivalent to that of a Corbino disk

(*Received* 24 *July* 1961; *in revised form* 21 *August* 1961)

IT IS well known[1] that for a given homogeneous and isotropic semiconductor, Corbino disk geometry provides the maximum magnetoresistance. The question arises: are there other geometrical arrangements that will provide equivalent magnetoresistance? The answer is yes. For example, it has been shown[1] that the magnetoresistance of a long rectangular wafer having electrodes placed along the larger dimension and the conduction path across the shorter dimension approaches that of a Corbino disk as the length-to-width ratio becomes very large. This, of course, is to be expected, for this is merely the condition for which cylindrical geometry and rectangular geometry become equivalent.

The conditions that are imposed on the other geometries having magnetoresistance equivalent to that of a Corbino disk are taken up in the following discussion.

The problem is, in essence, two-dimensional. Generally, the specimens are thin wafers (although

this is not a necessary requirement); the magnetic induction B is applied perpendicular to the surface of the wafer; and the current density J and the electric field E in the wafer lie in planes parallel to the surface. We assume perfectly conducting electrodes so that they *always* remain equipotential surfaces. In a specimen of finite dimensions, unless one of the electrodes is *circumscribed* by the other, or both the source and sink electrodes are circumscribed by a third floating electrode, space charge will always collect at the periphery of the wafer when a transverse magnetic field is applied, and hence will alter the equipotential lines. However, if the periphery is at infinity, or one electrode is circumscribed by the other, or both current electrodes are circumscribed by a floating third, *regardless of their shapes*, then the equipotential lines remain unaltered by the application of a magnetic field. For single-carrier conduction, the angle* between J and E is given by

$$\tan\Psi = \mu_B B \tag{1}$$

where μ_B is the Hall mobility in the transverse magnetic induction B. Hence, the component of current density parallel to E, i.e. parallel to grad V (V is the electrostatic potential), and the component of current density perpendicular to E, i.e. parallel to the equipotential lines, are given, respectively, by

$$J_E = |J| \cos\Psi = \sigma_B \cos^2\Psi |E| \tag{2}$$

$$J_V = |J| \sin\Psi = \sigma_B \sin\Psi \cos\Psi |E| \tag{3}$$

where σ_B is the physical conductivity of the wafer in transverse magnetic induction B (σ_0 is the conductivity for $B = 0$). The *effective* conductivity at every point of the semiconductor for the current flow along the E-lines (lines between source electrode and sink electrode) is given by the ratio J_E/E; hence, from (2),

$$\sigma\text{-(effective)} = J_E/E = \sigma_B \cos^2\Psi \tag{4}$$

If B is uniform throughout the specimen, then σ-(effective) will be constant throughout the specimen, and the ratio of the resistances (between the electrodes) of the specimen with and without applied magnetic field is simply

$$R_B/R_0 = \sigma_0/\sigma\text{-(effective)} = \sigma_0/\sigma_B \cos^2\Psi$$
$$= (\mu_0/\mu_B)(1+\tan^2\Psi) = (\mu_0/\mu_B)[1+(\mu_B B)^2] \tag{5}$$

In (5) we have set σ_0/σ_B equal† to μ_0/μ_B. This is valid so long as the magnetic field does not affect the carrier concentration, since then only the mobility is affected. The validity holds below the quantum limit.

Equation (5) is the formula for the magnetoresistance of a Corbino disk.[1] Thus, we have demonstrated conditions for which the total magnetoresistance is equivalent to that of a Corbino disk, namely: either one electrode must be inscribed within the other; or both current electrodes must be circumscribed by a third; or the periphery of the wafer must be at infinity, i.e. the spacing between the electrodes must be small in comparison with the distance of either to the periphery. The electrode shape is immaterial. Examples of the three types are demonstrated in Fig. 1: the single inscribed electrode in (a) (Corbino disk), (b), (c), (d) and (f); the double inscribed electrodes in (e) and (j); and the double electrode with the periphery at a great distance in (g) and (i). The configuration in (h) is a hybrid of the first and last types; the edges of the electrodes are remote from the periphery, and the parts near the periphery are self-circumscribed. The light lines of the diagrams represent the periphery of the specimen, and the dark lines, the circles, and the characters represent electrodes.

For two-carrier conduction, the effective conductivities of the carriers are additive, giving for the total effective conductivity

$$\sigma\text{-(effective)} = (\sigma_B)_n \cos^2\Psi_n + (\sigma_B)_p \cos^2\Psi_p \tag{6}$$

The n and p subscripts apply, respectively, to electrons and holes. The total magnetoresistivity is found from the second member of (5) by substituting (6) for σ-(effective).

* The Hall angle Ψ for a given material depends only on the local magnetic induction B. However, B is the resultant of the applied field and the field associated with the current through the specimen. For reasonable current densities, the effect of the latter can be neglected.[2]

† Strictly speaking, μ_0/μ_B in (5) should be the ratio of the drift mobilities. However, for isotropic material in general, the ratio for Hall mobilities should be equal to the ratio for drift mobilities.

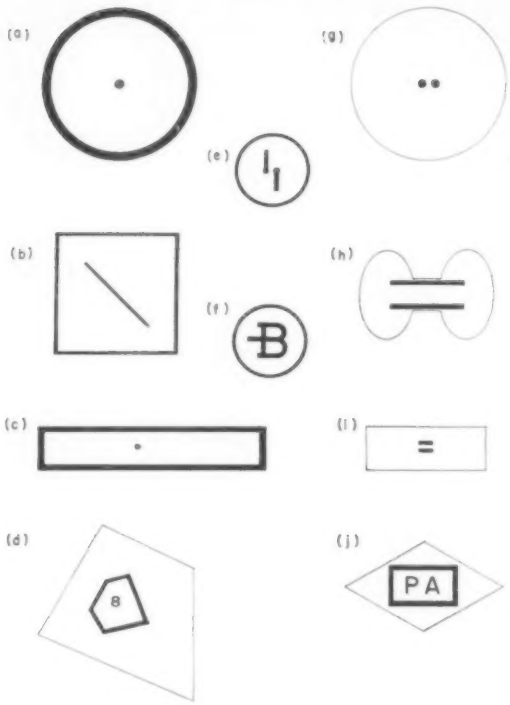


FIG. 1. Two-dimensional electrode configurations having magnetoresistance equal to the Corbino disk geometry, diagram (a), are shown in (b), (c), (d), (e), (f) and (j), and configurations that approximate the Corbino disk magnetoresistivity are shown in (g), (h) and (i). Heavy lines, dots, and characters are electrodes; light ones are the specimen periphery. Where three electrodes are shown, (e) and (j), the circumscribing electrode is floating.

The current flowing along the equipotential lines is a circulating one, since the equipotentials are closed curves (some of them close at infinity). By the use of the theory of complex variables, it can be shown[3] that, for single-carrier conduction, the ratio of the total circulating current, I_V, to the total current flowing between the electrodes, I_E, is

$$I_V/I_E = (d\tan\Psi)/(4\pi C) = (d\mu B)/(4\pi C) \tag{7}$$

where C is the capacitance, *in vacuo*, between the two electrodes, and d is the thickness of the specimen. For two-carrier conduction, (7) takes the form

$$\left(\frac{I_V}{I_E}\right)_p + \left(\frac{I_V}{I_E}\right)_n = \frac{dB}{4\pi C}(\mu_p + \mu_n) \tag{8}$$

In the above discussion, the electronic mobility μ_n is considered to be a *negative* quantity.

Burroughs Corporation MILTON GREEN
Burroughs Laboratories
Paoli, Pennsylvania

References

1. H. WEISS and H. WELKER, *Z. Phys.* **138,** 322 (1954).
2. M. GREEN, *J. Appl. Phys.* **32,** 1286 (1961).
3. W. R. SMYTHE, *Static and Dynamic Electricity* pp. 73–74. McGraw-Hill, New York (1939).

Growing heavily compensated germanium crystals of known impurity concentrations

(*Received* 15 *August* 1961)

IN EXPERIMENTAL work with impurities in germanium, it is often of interest to determine the effect of the impurities, or the effect of a specific pair of impurity types, on the electrical properties of the material. Since conventional measurements of conductivity and Hall coefficient can only be used directly* to determine the concentration of a single impurity type (N-type or P-type) and not the concentrations of the two impurity types present in heavily compensated samples, it is of particular interest to be able to process the germanium in such a way as to know the impurity concentrations. This is especially true if it is desired to keep one impurity type relatively constant while varying the other.

The floating crucible technique described previously[4] in conjunction with the normal CZOCHRALSKI[5] process has been used to obtain the desired control over the growing process so that a single majority impurity doping with a

* Several indirect methods are available but are open to question as to accuracy in heavily compensated samples due to incompleteness of our present understanding of the conduction process and the pertinent mobilities at low temperatures; e.g. Refs. 1–3.

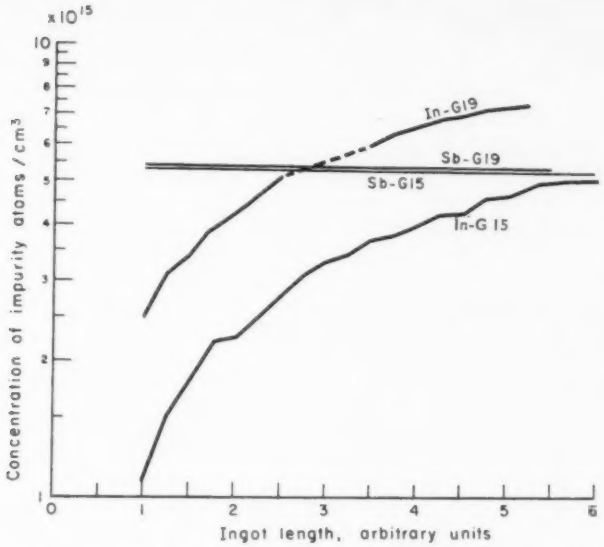


FIG. 1. Double doped germanium crystals G–15 and G–19 grown by the modified floating crucible technique.

varying degree of compensation with two impurities of nearly equal and small segregation coefficients can be achieved with a high degree of accuracy. Germanium containing a high concentration of one of the impurity types is used as a master alloy for the charge in the floating crucible. A particular weight of the master alloy corresponds to a given constant concentration of the impurity in the final grown crystal for a particular floating crucible size. Calibration of the crucible can be obtained with only the master alloy in the floating crucible and intrinsic germanium in the reservoir crucible. After calibration of the floating crucible, the crystal can be grown with the compensating impurity introduced into the reservoir crucible in any of the conventional ways. The compensation gradient can be readily determined from resistivity or Hall measurements since one impurity, that in the floating crucible, is practically constant and is known. Fig. 1 shows representative curves of crystals obtained in this manner.

Acknowledgements—The work reported herein was sponsored in part by the Electronics Research Directorate of the Air Force Cambridge Research Center under Contract AF 19(604)–8458. Appreciation is expressed to C. LEUNG and G. POTTS for growing the crystals.

Research Laboratories L. M. LAMBERT
Aeronutronic Division of
Ford Motor Company
Newport Beach, California

References

1. H. BROOKS, *Advanc. Electron. and Electron Phys.* **7,** 117–127 (1955).
2. D. LONG, C. D. MOTCHENBACHER and J. MYERS, *J. Appl. Phys.* **30,** 353 (1959).
3. P. A. LEE, *Brit. J. Appl. Phys.* **8,** 340 (1957).
4. W. F. LEVERTON, *J. Appl. Phys.* **29,** 1241 (1958).
5. J. CZOCHRALSKI, *Z. Phys. Chem.* **92,** 219 (1918).

Growth of semiconducting compounds from non-stoichiometric melts

(*Received* 5 *September* 1961)

IT IS possible to grow semiconducting compounds from non-stoichiometric melts. In fact it is sometimes desirable to do so where it is technically difficult to produce stoichiometric melts; for

example, for those III–V compounds for which the equilibrium vapour pressure of the Group V element at the melting point is large (e.g. the equilibrium vapour pressure of phosphorus at the melting points of GaP and InP are respectively > 20 atm[1] and approximately 60 atm[2]).

The purpose of this letter is to describe the characteristics of growth from a non-stoichiometric melt and to show that they can be considered from the viewpoint of constitutional supercooling. As an illustration, we describe the inhomogeneities which exist in an indium antimonide single crystal which has been grown from an indium-rich melt under conditions of constitutional supercooling and indicate how these inhomogeneities can be prevented by a choice of growth conditions which avoid constitutional supercooling. Theoretical work predicting the onset of constitutional supercooling during growth from a simple solute/solvent binary melt using the Czochralski rotating crystal technique has been published by HURLE[3]. Experimental work validating the theory for the case of gallium doped germanium has been carried out by BARDSLEY *et al.*[4].

The growth of compound semiconductors from non-stoichiometric melts can be considered in relation to the phase diagram shown in Fig. 1. The compound (AB), for example a III–V compound, is formed at 50 at. per cent A and eutectics occur at nearly pure A and at nearly pure B. If growth is proceeding into a melt having, for example, an excess concentration of component A, then the equilibrium liquidus temperature of the melt (T_A) will be depressed by ($T_m - T_A$) below the equilibrium temperature of the stoichiometric melt (T_m). The excess concentration of component A will be rejected into the melt at the solid–liquid interface and the A-rich boundary layer so formed can give rise to a zone of constitutional supercooling.

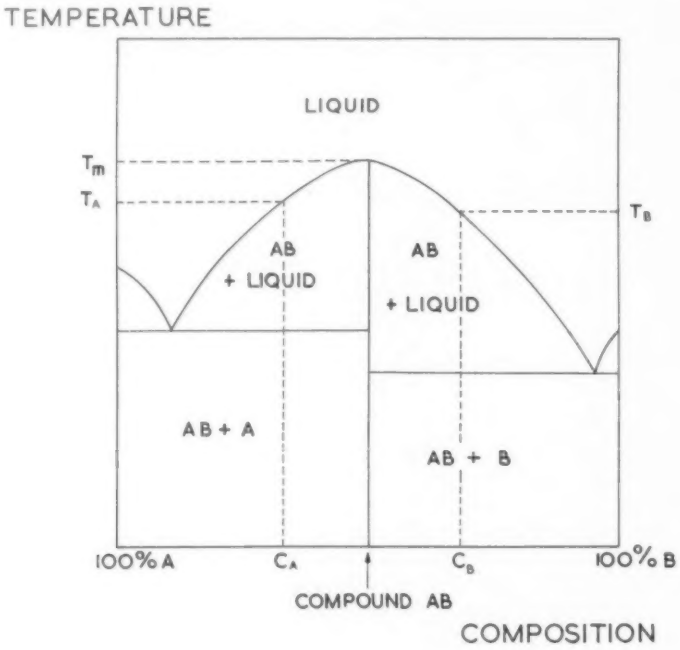


FIG. 1. Phase diagram for a binary system showing compound formation at composition AB.

The magnitude of the gradient of constitutional supercooling in the melt at the interface $(dS/dx)_{x=0}$

for the case of growth from a stirred melt of a solvent containing a solute is[3]

$$\left(\frac{dS}{dx}\right)_{x=0} = -\frac{mfC_L(1-k)}{D\{k+(1-k)\exp(-\Delta)\}} - \frac{K_sG_s - L\rho f}{K_L} \quad (1)$$

where:

m = slope of the liquidus line at the appropriate melt concentration from the phase diagram for the binary alloy
f = growth rate
D = diffusion coefficient for the solute in the molten solution
k = distribution coefficient
C_L = concentration of solute in the molten solution
K_s, K_L = thermal conductivities of the solid and liquid, respectively, at the equilibrium temperature
L = latent heat fusion of the solution
ρ = density of the solution
Δ = $f\delta/D$ where $\delta = 1{\cdot}6D^{1/3}\nu^{1/6}\omega^{-1/2}$
ν = kinematic viscosity of the melt
ω = angular velocity of the rotating crystal.

Equation (1) can be applied to the growth of the compound AB from a non-stoichiometric melt as follows: the distribution coefficient (k) for excess A or B in AB is either extremely small or zero for the case of a III–V compound* and hence equation (1) can be approximated to:

$$\left(\frac{dS}{dx}\right)_{x=0} = -\frac{mfC_L\exp(\Delta)}{D} - \frac{K_sG_s - L\rho f}{K_L} \quad (2)$$

The concentration C_L is to be interpreted as the excess concentration of A (or B) in the melt, i.e. $C_L = C_A - 50$ at. per cent.

On the basis of the work on degenerately doped germanium,[4,5] it might be expected that the existence of a small gradient of constitutional supercooling during the growth of a semiconducting compound from a non-stoichiometric melt should cause the solid–liquid interface to adopt a cellular form. To test this expectation, we have grown, using the Czochralski technique, a [100] oriented single crystal of indium antimonide from an indium-rich melt under conditions such as to produce a zone of constitutional supercooling in the melt ahead of the solid–liquid interface. Fig. 2(a) shows the solid–liquid interface of the crystal, preserved by rapidly withdrawing the crystal from the melt during growth. The interface is cellular and very similar to interfaces obtained on [100] oriented gallium-doped germanium crystals.[5]

Using an optical goniometer, we have found that the faces of the cells shown in Fig. 2(a) are of the type {111}. BARDSLEY *et al.*[4] found that the faces of the cells on the interfaces of germanium single crystals were inclined at only a small angle to the plane of the interface and were not {111} planes. However, they subsequently showed[5] that, during growth, the faces were {111} planes, and that the reason that these faces did not appear on preserved interfaces was that melt adhered in the boundaries between the cells during the rapid withdrawal of the crystal from the melt. The melt from which the crystal shown in Fig. 2 was grown was also lightly doped with radio-tellurium (Te concentration 15 p.p.m.); Fig. 2(b) is an autoradiograph of a section cut and ground parallel to and just above the plane of the preserved solid–liquid interface. The light areas are tellurium-rich and delineate the faces of the cells. The dark lines which separate the light areas delineate the boundaries between the cells. Fig. 2(c) is a photograph, taken in dark field illumination of the same surface after etching in CP_4 etchant. The enhanced tellurium concentration on the cell faces [Fig. (2b)] further proves that the cell faces are planar {111} facets; the excess tellurium concentration is a consequence of the large "facet effect" exhibited by tellurium in indium antimonide.[8,9] Thus a cellular structure with faceted {111} surfaces has been produced by growing under conditions of constitutional supercooling from an indium-rich melt; the faceted surfaces are delineated by the radio-tellurium which is preferentially incorporated on them. Similar structures have been

* For many of the other semiconducting compounds however, (for example lead selenide[6], lead telluride[7]) finite solubility of A (or B) in the compound AB has been observed.

produced by growing from an antimony-rich melt.

A feature of the cellular structure is that, for distribution coefficients $k < 1$, solute is trapped in the depressed boundaries between the cells.[10] Under extreme gradients of constitutional supercooling the solute (in the case considered, the excess of component A) can form a separate phase at the cell boundaries. In the case of indium antimonide, the indium or antimony solid phase will be formed when the indium or antimony concentration in the melt at the cell boundaries reaches the InSb–In or InSb–Sb eutectic compositions[11] respectively. Similarly, the precipitation of metallic gallium in gallium phosphide grown from the melt at low phosphorus pressures has been experienced in this laboratory and elsewhere. If the composition of the melt is known, a consideration of equation (2) permits the choice of a growth rate which will be sufficiently slow to avoid a gradient of constitutional supercooling [i.e. $(dS/dx)_{x=0} < 0$] and hence the cellular structure and the precipitation of the second phase.

Acknowledgements—It is a pleasure to acknowledge the experimental assistance of Miss V. Doyle. This letter is published by permission of the Controller, H.B.M. Stationery Office.

Ministry of Aviation
Royal Radar Establishment
Malvern, Worcestershire
England

D. T. J. Hurle
O. Jones
J. B. Mullin

References

1. C. J. Frosch and L. Derrick, *J. Electrochem. Soc.* **108,** 251 (1961).
2. J. van den Boomgaard and K. Schol, *Phillips Res. Rep.* **12,** 127 (1957).
3. D. T. J. Hurle, *Solid-State Electron.* **3,** 37 (1961).
4. W. Bardsley, J. M. Callan, H. A. Chedzey and D. T. J. Hurle, *Solid-State Electron.* **3,** 142 (1961).
5. W. Bardsley, J. M. Callan, H. A. Chedzey and D. T. J. Hurle, to be published.
6. A. E. Goldberg and G. R. Mitchell, J. *Chem. Phys.* **22,** 220 (1954).
7. E. Miller, K. Komarek and I. Cadoff, *Trans. A.I.M.E.* **215,** 882 (1959).
8. K. F. Hulme and J. B. Mullin, *Phil. Mag.* **4,** 1286 (1959).
9. J. B. Mullin and K. F. Hulme, *J. Phys. Chem. Solids* **17,** 1 (1960).
10. J. W. Rutter and B. Chalmers, *Canad. J. Phys.* **31,** 15 (1953).
11. T. S. Liu and E. A. Peretti, *Trans. Amer. Soc. Metals* **44,** 539 (1952).

Current gain in metal–insulator tunnel triodes

(*Received* 31 *October* 1961)

Spratt *et al.* have described a three-terminal device consisting of an n-type Ge substrate with successive layers of Al, Al_2O_3 and Au deposited upon it.[1] They find that this structure exhibits amplifying characteristics similar to those of an n–p–n-transistor, with current gains typically in the range 0·6–0·8, and sometimes as high as 0·97, but never greater than unity.[2] They interpret their observations as evidence that electrons tunnel from the Au to states well above the Fermi level in the Al, and that these hot electrons are able to pass through the Al film with sufficient energy remaining that they can surmount the potential hill at the Ge interface and be collected. In the following, we discuss several factors which can be expected to limit the current gain in such a tunnel triode. We conclude that tunneling cannot account for the observed behavior, and offer an alternative mechanism of operation.

Scattering within the Al layer or in the Ge collector will limit the current gain. Electron–electron scattering in metals has been treated by Wolff.[3] Estimates based upon his work lead to mean-free-paths of 30–60 Å for electrons having a few eV of energy in typical metals, including Al. Wolff's equation (20) is valid in the energy range $E_F < E < 2E_F$ of interest here, if the factor $(1-7E_F/5E)$ is replaced by

$$[1-7E_F/5E+2E_F(2-E/E_F)^{5/2}/5E].$$

For our purposes, it is important to note that his analysis is concerned specifically with collisions in which the primary electron loses a substantial fraction (e.g. 1/2) of its energy. Escape depths of this same magnitude are deduced from secondary electron emission from Pt and Ge,[4,5] and Ni[6]. An isolated case where larger escape depths (1000 Å) have been reported is for photoelectron emission from potassium.[7] On the other hand, a different interpretation of these data, consistent with a small escape depth, has been suggested[8]

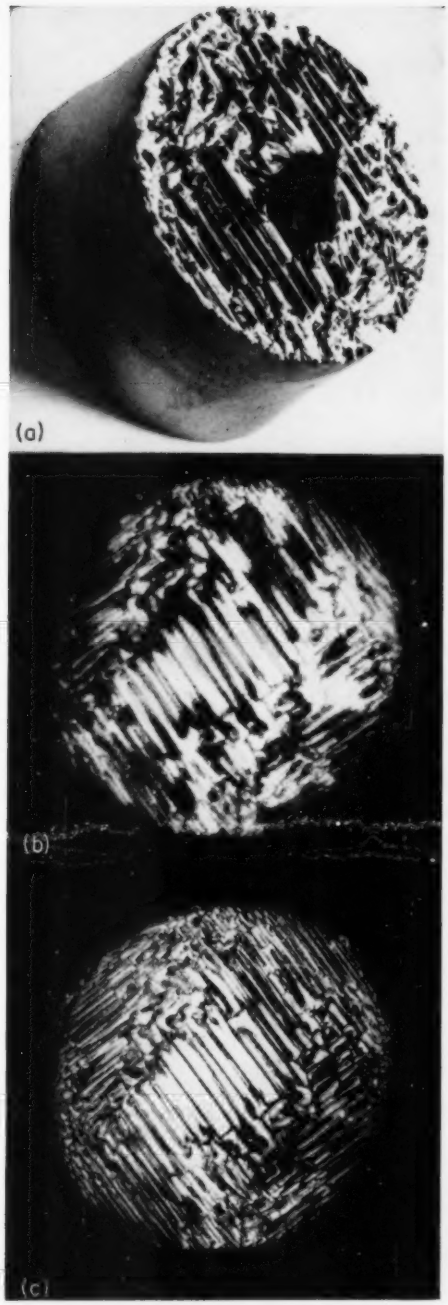


FIG. 2. Cellular structure in a [100] oriented single crystal of InSb grown from an In-rich melt (a) (Top) preserved interface showing cellular structure composed of {111} facets. (b) (Centre) auto-radiograph of transverse section in cellular region showing enhanced concentration of radio-Te on {111} facets. (c) (Bottom) photomacrograph of section shown in (2b) after etching.

[*facing p.* 320

In the case of Al, front and back surface photoelectron emission studies of thin evaporated films have shown that the escape depth is less than the absorption length (70 Å) of the light used, but do not indicate how much shorter it might be.[9] From this evidence, one would expect the current gain to be limited to values of the order of 0·2 due to scattering in the 80 Å Al layer alone.

Furthermore, an electron which has traversed the Al layer and entered the space-charge region of the Ge collector has a high probability of being scattered back into the Al again before its energy drops below the potential maximum at the metal–semiconductor interface. The mean-free-path for scattering of hot electrons by optical phonons in Si and Ge is believed to be approximately 50 Å,[10] a value which is many times smaller than the space-charge layer thickness of the collector junction. This is the dominant energy loss mechanism and gives rise to large-angle scattering with 0·038 eV of energy lost per collision. A typical electron will enter the conduction band of the Ge with an energy many times greater than that of the optical phonon, and must suffer a corresponding number of large-angle collisions before its destiny as a "collected" electron is established. The fraction of back-scattered electrons may be calculated from the closely related Albedo problem[11] which slightly under-estimates reflection in the present problem and gives reflection coefficients of 0·46 and 0·59 for 10 and 20 collisions respectively.

Reflection of the electrons at the Al–Ge interface will further reduce the current gain. The propagation vector describing an electron wave of given energy must change discontinuously at this interface, resulting in partial reflection of the incident wave. One would expect that reflection at the interface between a microcrystalline Al film and Ge would be quite appreciable.

Each of the processes discussed above are expected to produce losses in electron transfer that are an order of magnitude greater than the minimum (3 per cent) reported for such triodes. We have ignored losses due to the presence of the oxide layer which is almost certainly present at the Ge surface, to scattering by phonons and imperfections in the Al film, and to the fact that not all of the tunneling electrons will have energies sufficient to surmount the potential hill at the Ge surface barrier. For these reasons we conclude that the behavior of this device cannot be explained by the proposed mechanism.

We suggest instead that the Au film contacts the space-charge region within the Ge through the pinholes that must almost certainly be present in the Al film, as shown in Fig. 1, giving rise to a

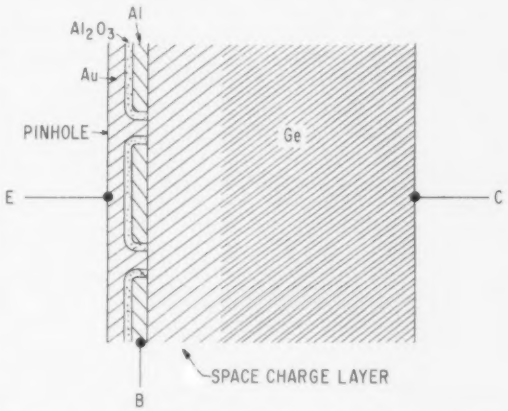


FIG. 1. Metal interface amplifier with pinholes in base layer.

depletion layer transistor[12] in which electrons are injected by the Au emitter directly into the space-charge layer formed between the Al and the Ge. This emitter contact is separated from the Al boundary of the space-charge layer only by the thickness of the Al_2O_3 layer, which is many times less than that of the space-charge region. Amplification is thus similar to that occurring in an *n–p–n*-transistor except that in the latter the injected electrons must first diffuse through the base before reaching the space-charge region. This model readily accounts for the similarity of the electrical characteristics to those of an *n–p–n*-transistor and for the observed influence of the collector voltage upon the emitter current. The latter is evidence that pinholes are indeed present in the Philco devices.

Acknowledgment—These investigations were supported by the Air Force Cambridge Research Laboratory under Contract No. AF 19(604)–6623.

General Electric Research Laboratory R. N. HALL
Schenectady, N.Y.

References

1. J. P. SPRATT, R. F. SCHWARZ and W. M. KANE, *Phys. Rev. Letters* **6,** 341 (1961).
2. J. P. SPRATT, R. F. SCHWARZ and W. M. KANE, *Solid-State Device Research Conference. Trans. I.R.E.*, abstract to be published.
3. P. A. WOLFF, *Phys. Rev.* **95,** 56 (1954).
4. A. J. DEKKER, *Solid State Phys.* **6,** 251 (1958).
5. J. B. JOHNSON and K. G. MCKAY, *Phys. Rev.* **93,** 668 (1954).
6. H. BRUINING, *Physics and Applications of Secondary Electron Emission*, p. 100. McGraw-Hill, New York (1954).
7. H. THOMAS, *Z. Phys.* **147,** 395 (1957).
8. H. SIMON and R. SUHRMANN, *Der lichtelektrische Effekt und seine Anwendungen*, pp. 65, 88. Springer, Berlin (1958); G. A. KATRICH and O. G. SARBEI, *Solid-State Phys.* **3,** 1629 (1961).
9. E. A. TAFT, General Electric Research Laboratory. Unpublished data.
10. J. L. MOLL, N. I. MEYER and D. J. BARTELINK, *Phys. Rev. Letters* **7,** 87 (1961).
11. G. PLACZEK, *The Science and Engineering of Nuclear Power*, (Edited by C. GOODMAN) Vol. 2, Chap. 7. Addison–Wesley Press, Cambridge, Mass. (1949).
12. W. W. GAERTNER, *Proc. I.R.E.* **45,** 1392 (1957).

Solid-State Electronics Pergamon Press 1961. Vol. 3, pp. 323–326. Printed in Great Britain

BOOK REVIEWS

F. P. BUNDY, W. R. HIBBARD, JR. and H. M. STRONG: **Progress in Very High Pressure Research.** Wiley, New York, 1961. 314+xix pp., $12.00.

THIS book contains the proceedings of an international conference on high pressure sponsored jointly by the Air Research and Development Command and General Electric Research Laboratory at Lake George, New York in June 1960. The text consists almost completely of a reproduction of the papers presented at the conference in substantially the form submitted by the authors before the beginning of the conference, plus selected comments transcribed from tape recordings made during the discussion periods which followed each paper.

The Preface says the book was prepared as rapidly as possible, with a minimum of editing in order that it may be available to new workers in the field as well as the older ones. An identical version save for paper cover, the discussion after each paper, and the half-dozen pages inserted later, was distributed at the beginning of the conference. It is regrettable that the hard-cover version reviewed here appeared a year and a week after the conference edition was distributed. Both versions are in offset, apparently from typewritten copy. Added to the later edition are the transcribed comments mentioned, an English version of a paper delivered in Russian, a later writeup of an important disclosure made by KENNEDY, and some prepared discussions written by DACHILLE. Some errors such as "unattached" and "cyrstal" instead of "unattacked" and "crystal" slipped through the retyping. The fact that the book costs $12.00 must I suppose be attributed to the potentially small sales because of the esoteric contents and the fact that the Air Force cannot subsidize the whole scientific public, rather than to a great deal of labor other than by secretaries. Conferees seem to have received both versions as part of having attended.

The papers divide roughly into two groups—technique with little discussion of results and results with little discussion of technique. Implicit in some of the technique discussion is the assumption that ultrahigh pressures have been refined away from being primarily useful for raw synthesis or qualitative observation to a stage at which precise measurements of such variables as pressure and temperature have considerable meaning. G. C. KENNEDY punctured this illusion by replacing his planned talk on the SiO_2–H_2O system with work he had just done on calibration of the phase transitions used for constructing a high-pressure scale.

The gist of his talk, and the paper subsequently published here, shows that BRIDGMAN some years before had obtained two sets of pressures above 25 kilobars at which the metals now used for calibration underwent transitions—a lower set determined under relatively good hydrostatic conditions and a higher set obtained by compression in the presence of a great deal of friction. BRIDGMAN had taken pains to point out the conditions under which each set of points had been obtained, but apparently most workers simply adopted the higher set of values for calibration. The result was that any pressure quoted above 25 kilobars in the pre-conference literature, or in the papers of this book, is roughly thirty per cent too high. For this reason, KENNEDY's paper should be read first by any newcomer.

Apart from this, the book is quite interesting because the non-technique papers illustrate what is being done at very high pressure, and not so high pressure, in several different areas, and because it gives a glimpse of what the Russians are doing with the new techniques. A large number of papers deal with phase transformation under pressure while the topic receiving the second greatest amount of space is synthesis; the remainder of the book is given over primarily to diffusion, elasticity, low temperature experiments, semiconductors, and magnetic effects under pressure. Apart from a greater interest in metal forming with pressure, the Russians appear to be doing a very competent job along much the same lines as Western workers. The articles also

reveal what is not being done—examination of whether any real hydrostaticity prevails, whether pressure gradients affect results, and learning to analyze simultaneous application of pressure and temperature, for example. The book is not as much an introduction to the art as a survey of what its title promises—progress in the field.

DOUGLAS WARSCHAUER

J. THEWLIS (Editor-in-Chief): **Encyclopaedic Dictionary of Physics.** Volume 1: **A–Compensated bars.** Pergamon Press, Oxford, 1961. £98 ($298.00) per set of eight volumes.

THIS is an extremely important work: no comparable dictionary of physics has appeared in English since Glazebrook's *Dictionary of Applied Physics* published in 1923, although several smaller ones have appeared recently. The complete *Dictionary* will consist of 8–10 volumes each of about 800 pages, page size 25×19 cm. A seven language (English, French, German, Russian, Spanish, Chinese, Japanese) Glossary of 15,000 physical terms will be included, as well as a Subject Index volume. The Editor-in-Chief and Associate Editors, as well as the Publishers, must be warmly congratulated on the appearance of this first product of their truly Herculean task, and it is to be hoped that the other volumes will follow in rapid succession.

The *Dictionary* is arranged in strict alphabetical order and covers to a greater or less extent, besides conventional pure and applied physics, topics in Mathematics, Chemistry, Electrical Engineering, etc. of direct importance to physics.

The Editor has had the benefit of the help of an impressive array of Consulting Editors (British and American). The contributors, at least to this first volume, appear to have been chosen with care as most likely to write well informed and up-to-date contributions. With few exceptions, no article is longer than 2000–3000 words, each being intended to be self-contained, though cross-references are given, even if not as plentifully as one might have desired. The general aim has been not to write articles for specialists, but to keep throughout a near graduate standard. Inevitably, there is some unevenness between the various contributions, but quite a number of articles are a model of how to convey a wealth of information with clarity and conciseness: one already sees in one's imagination the innumerable text-books that will reproduce them . . . It is only with regret that one abstains from listing those contributions which to one particular physicist have appeared most outstanding.

A work of this kind is of course intended both for physicists who wish to collect information rapidly and painlessly, and for chemists, biologists, etc. who only too often need the help of physics in their own research. To test the success of the *Dictionary* within this wider circle, three kind colleagues (a chemist, a botanist, a zoologist) agreed to act as guinea-pigs reading selected articles which interested them. Their verdict was extremely favourable, the clarity of the exposition being unanimously stressed. The chemist, as one might perhaps have expected, criticized the choice of the chemical items to be covered and also some entries in which he is a specialist. Understandably, there will be a tendency in some physicists to take up a similar attitude, but the general impression of the Reviewer is that the choice of entries and the space allotted to them reveals, from everybody concerned, prolonged and intelligent thought and that, given the aim and size of the work, it would have been most difficult to do better. The only adverse comment is that sometimes the final editing has not been thorough enough. For instance, there are repetitions, e.g. under "bolometer, superconducting" the definition of bolometer is repeated, although given in the previous entry. In other cases, on the other hand, the main definition is missing, e.g. in "anomalous dispersion" (it will undoubtedly come later under "dispersion"): this short entry refers somewhat surprisingly to the old Julius theory of solar phenomena. As for entries that one would have expected, but do not appear in the main alphabetical sequence, for instance "band theory", "band spectra", "band width" or "avalanche phenomena in semiconductors and dielectrics", no doubt their location will be given in the Index. Also, some entries are tautologic, or appear hardly necessary, e.g. "barrier-layer capacitance: the effective capacitance of the barrier-layer formed at the boundary between two conductors". Of course it would be easy, and most unfair, to pick up small blemishes of this kind.

The production is generally of a high quality. However, a few mistakes have been noticed, particularly in proper names. For instance, in the articles on barrier-layer rectification, a certain "SCHOTTLEY" appears, possibly a cross between SCHOTTKY and SHOCKLEY, and BARDEEN becomes BARDUN. Particularly unfortunate is that the very first word in the *Dictionary* is wrong, the German physicist Ernst Abbe being given an accent thereby changing into a French abbot.

Just to give solid-state physicists an idea of the scope of the work, under "acceptor" we have "acceptor atom" (12 lines) with its properties defined according to the band model; then a longer article on "acceptor centre", with definitions of its ionization energy, occupation probability (with formula), a Table of the ionization energies of impurities in germanium and silicon and the physical picture of centres due to substitutional impurities, followed by four cross-references (for instance to "Activation Energy", where additional information is provided) and a bibliography of four entries; finally, "acceptor level" (9 lines) mentions the variation of activation energy with impurity concentration and the distortion of the energy levels near the surface of the crystal. "Acceptor" does not appear however in the sense of a circuit which accepts a certain frequency.

Clearly, with a book of this kind, the review can only deal with very few examples. But it should not be closed without trying to convey the general impression that here is an outstanding achievement that will be profitably consulted for many years and which physicists will like to peruse from time to time to be instructed, reminded and entertained. The Editors deserve admiration and gratitude.

L. PINCHERLE

G. W. A. DUMMER and J. M. ROBERTSON (Editors): **British Miniature Electronic Components and Assemblies Data Annual 1961–62**. Pergamon Press, Oxford, 1961. xviii + 479 pp., 80*s*.

THIS is not a technical book in the text book sense, but a collection of data on components of British manufacture which should greatly ease the job of the circuit designer, particularly if using transistors (though transistors themselves are specifically excluded). "Miniature" is defined as less than 1 in. cube though items over this limit are included where they are the smallest of their particular type. Sub-assemblies are also included (for example, miniature transistorized amplifiers) which establishes a precedent for including microminiature and solid circuit assemblies in the future.

On the other hand the book is more than just a catalogue of various manufacturers' products. Some very valuable sifting has been done—and two better qualified sifters would be hard to find. The book replaces a complete catalogue library as far as miniature components are concerned and as such should be extremely useful to circuit designers, none of whom have enough time to read all the available catalogue information.

To keep it up to date, the book is to be published annually. At the present price—which seems excessive in view of the advertising value—the private individual is unlikely to place a regular order. However, no design group should be without the current issue on the laboratory shelf. Referring to prices brings me to my only criticism of an otherwise invaluable volume: that the price of the products described is not mentioned. The Editors say that prices are omitted as they are subject to change, but surely no large changes are likely to occur in twelve months and every circuit designer is, or should be, price conscious.

Finally, I should like to congratulate the Editors on what must have been a Herculean task well done. That they plan similar annual volumes on American and European components, plus an annual volume on British transistors and sub-miniature valves can only mean that they are gluttons for punishment.

ALAN F. GIBSON

C. K. MOORE and K. J. SPENCER: **Electronics: a Bibliographical Guide.** Macdonald, London, 1961. xvii + 411 pp., 65*s*.

IT IS a commonplace that a major problem facing the scientist is that of apprehending the ever increasing volume of scientific literature. The Publishers describe this volume as a "new approach to literature searching". The main body of the book is split into sixty-seven specialist subject sections in which are listed books, bibliographies,

survey and review papers, published conference proceedings and a variety of papers selected as outstanding in some respect, e.g. the paper's pioneering nature or historical interest. Under this last heading there is endless room for disagreement with the selection made. Each entry is accompanied by a brief note on the number of references quoted, the topic or topics covered and, in the case of a book, the relevant chapter headings with some indication of the type of treatment. There are about 3000 such entries, covering the period from 1945 to June 1959. In addition, there is a list of nearly 300 "sources of information", e.g. bibliographies of bibliographies, abstracting services and data books.

The reader is thus presented with signposts to the literature of each specialized field and with references which should allow him to pursue any particular topic. Specifically, the solid state is covered by the sections on phosphors and luminescence, infra-red, dielectrics, piezoelectricity, super- and photo-conductivity, semiconductor physics, magnetic materials, ferrites and paramagnetic resonance. Applications of solid-state devices are covered in the sections on masers, rectifiers, delay lines, transistors and magnetic, dielectric and parametric amplifiers.

As far as I can judge, the selection of items has been made intelligently and the compilers have used to good advantage the experience gained from the scientist–librarian relationship which has developed in their respective libraries. I have noticed no serious omissions though some of the references included surprise me. There are no doubts about the difficulties of literature searching and equally no doubt that this volume will be invaluable to any individual or group launching out into a new field of endeavour. I wonder, however, if too big a subject has been tackled and hence too thinly. Anybody with two or three years experience in one of the specialist fields will immediately recognize most of the references listed under that heading. How often does an individual move from one branch of electronics to another? In short, I see this book as a must for the library of any laboratory or institution of reasonable size but unlikely to be bought by the private individual. Furthermore, supplementary editions will be essential to keep it up to date—and it would be a pity not to build on such a solid foundation, the building of which must have been an enormous labour.

ALAN F. GIBSON

ERRATUM

M. GREEN, I. N. GREENBERG, J. BANZHAF, A. SASS, H. S. WHITE and R. V. MCKNIGHT: Electrical characteristics of a reduced-area, rhodium-plated, silicon end contact. Vol. 3, No. 1, pp. 1–10.

Line 4 of the footnote on p. 10 should read "(small area contact positive)" instead of "(small-area contact)".

Ref. 5, p. 10, should merely read "To be published.".

The sixth line from the bottom of p. 9 should read "Fig. 12 (a–c)" instead of "Fig. 12 (a–e)".

3
1